Universidade de Brasília
Departamento de Estatística

Modelo de regressão beta inflacionado em zero e um: uma aplicação à proporção de mulheres nas empresas

Jean Sabino Magalhães Diniz
Daniel Leite Martins Melo

Relatório apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2018

Jean Sabino Magalhães Diniz
Daniel Leite Martins Melo

Modelo de regressão beta inflacionado em zero e um: uma aplicação à proporção de mulheres nas empresas

Orientador:
Prof. Dr. **Leandro Tavares Correia**

Relatório Parcial apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2018

Resumo

Modelos de regressão beta são comumente utilizados para modelar proporções no intervalo contínuo. Entretanto, esses modelos não são convenientes na presença de 0 e/ou 1. Para ajustar problemas nos quais a variável dependente está no intervalo [0,1], abordamos neste trabalho a distribuição beta inflacionado em zero e um, que consiste na mistura da distribuição beta e de Bernoulli, e posteriormente a regressão beta inflacionado em zero e um; discutimos nesse trabalho aspectos inferenciais do modelo e também sua interpretação. Como aplicação para o tema, utilizamos esse modelo para verificar os fatores que influenciam a proporção de mulheres nas empresas. A escolha desses dados é de mera magnitude devido ao quadro social brasileiro. Podemos notar através desses dados a presença de empresas no qual não havia nenhuma mulher, isto é, a proporção é igual a 0, e empresas que só haviam mulheres, proporção igual a 1. Por fim, após feita a seleção e análise de diagnóstico do modelo, identificamos e interpretamos os fatores que influenciam a proporção de mulheres nas empresas.

Palavras-chave: Proporção, Distribuição beta inflacionado em zero e um, Regressão beta inflacionado em zero e um, Proporção de mulheres nas empresas

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Na literatura, existe uma enorme variedade de métodos estatísticos que podem ser utilizados para modelar dados. Para tal, é importante compreender os diferentes tipos de modelos propostos para que a análise da relação entre a variável resposta e as demais covariáveis seja feita da melhor maneira possível.

Devido a funcionalidade e com os avanços tecnológicos, estudos que envolvem modelos de regressão sempre estiveram em constante desenvolvimento e foi percebido que para alguns tipos de variáveis resposta a atribuição da distribuição normal não era razoável, influenciando o desenvolvimento de técnicas que fossem adequadas para o uso de outras distribuições assumidas pela variável resposta. Assim foram propostos modelos de regressão que utilizassem outras distribuições de probabilidade, por exemplo as distribuições de Poisson, binomial, gama, e normal inversa. Tendo como base as propriedades de suficiência estatística, foram agrupadas estas e outras distribuições de probabilidade, numa classe denominada como família exponencial unidimensional com dispersão. Esta classe de distribuições, serviu de base para que fossem elaborados os Modelos Lineares Generalizados (MLG) (NELDER; WEDDERBURN, 1972).

Dados com valores entre zero e um são de ocorrência comum na no dia-a-dia, como dados clínicos, notas de avaliações escolares, dentre outros. Para essas situações foi proposta a classe de modelos de regressão beta (FERRARI; CRIBARI-NETO, 2004), onde a variável resposta possui distribuição beta. Esses modelos são aplicados para entender as características específicas de inúmeros fenômenos, como por exemplo, analisar a eficiência dos municípios paraibanos em relação aos recursos do programa bolsa família (LOPES, 2017). Ou seja, a distribuição beta é comumente utilizada para modelarmos proporções, ou modelagem de objetos que pertencem ao intervalo $(0, 1)$, uma vez que, essa distribuição está definida nesse intervalo.

Em alguns casos, no entanto, os dados podem apresentar ponto de massa tanto em zero quanto em um. Como por exemplo, a proporção de enfermeiros com curso superior nos municípios brasileiros e a proporção de óbitos em menores de 1 ano por causa mal definidas nos municípios brasileiros do ano 2000 descrita por Ospina (2008). Desta forma surge como suporte o intervalo [0,1], mas a distribuição beta não admite os valores zero e um em seu suporte. Para esses tipos de dados é interessante uma estrutura de probabilidade para os pontos de massa e as realizações contínuas, ou seja, é possível e desejável utilizar uma mistura entre a distribuição discreta que modela os valores 0 e 1 e a distribuição contínua para os demais valores.

A distribuição BIZU (OSPINA, 2008), que utiliza a ideia de mistura de distribuições (MCLACHLAN; PEEL, 2004), é uma opção teórica para esse tipo de dados. Nela o

componente discreto é modelado pela distribuição de Bernoulli e o componente contínuo é modelado pela distribuição beta,.

O termo inflacionado denota que a massa de probabilidade de alguns pontos é maior do que a presumida no modelo em questão. A maior parte das análises na área de modelos inflacionados deduz que a distribuição da variável resposta é uma mistura entre uma distribuição degenerada no zero e uma outra distribuição conhecida.

Na literatura, mais especificamente em (OSPINA, 2008), encontramos proposto o modelo de regressão beta inflacionado em zero e um. O autor versa sobre uma extensão do modelo de regressão beta para casos em que a variável resposta varia no intervalo [0, 1]. O modelo admite que a distribuição da variável resposta é beta inflacionada, sendo esta uma mistura de uma distribuição beta e uma distribuição de Bernoulli. Nesse modelo, tanto a média dos valores contínuos como a probabilidade de assumir os valores zero e um são adaptadas em função de variáveis preditoras.

Este trabalho é motivado por um conjunto de dados concedido pela fundação Instituto de Pesquisa Econômica Aplicada (Ipea), proveniente da Relação Anual de Informações Sociais (RAIS) de 2016 (ME, 2015). O Ipea tem por finalidade realizar pesquisas e estudos sociais e econômicos, dando apoio técnico e institucional ao Estado brasileiro na avaliação, formulação e acompanhamento de políticas públicas e programas de desenvolvimento.

Portanto, temos como objetivo para esse trabalho propor um modelo inferencial para a proporção de mulheres nas empresas através de um modelo de regressão beta inflacionado em zero e um. De maneira mais especifica estudar mistura de distribuições; estudar o modelo de regressão beta inflacionado em zero e um; aplicar o modelo estudado no conjunto de dados disponibilizado pelo Ipea; e avaliar a participação do sexo feminino na economia brasileira.

2 Revisão de Literatura

Neste capítulo será aludido conceitos e definições com o objetivo de tornar mais fácil e claro a compreensão dos métodos que serão utilizados no desenvolver deste trabalho. Primeiramente, será discutido sobre o método da máxima verossimilhança, que será o principal método de estimação; posteriormente, será discutido sobre as principais distribuições que serão utilizadas; e em seguida, será discutidos sobre modelos e análise de diagnóstico. O foco maior deste capítulo é falar sobre a distribuição Beta inflacionada em zero e um (Seção 2.3) e sobre o modelo de regressão Beta inflacionada em zero e um (Seção 2.7).

2.1 Método de Máxima Verossimilhança

Pode-se dizer que em estatística um dos métodos mais usados para se obter estimadores é o método da máxima verossimilhança. Considerando uma amostra aleatória $Y_1, Y_2, ..., Y_n$ cada uma com função de densidade ou probabilidade $f(y_t|\theta)$, com $\theta = (\theta_1, \theta_2, ..., \theta_r) \in \Theta$, onde Θ é o espaço paramétrico. A função de verossimilhança de θ é dada por por:

$$L(\theta; y_t) = \prod_{i=1}^{n} f(y_t|\theta).$$

O estimador de máxima verossimilhança de θ é o valor $\hat{\theta} \in \Theta$ que maximiza a função de verossimilhança $L(\theta; y_t)$. Contudo, é fácil verificar que, satisfeita as condições de regularidade, encontrar o valor que maximiza o logaritmo natural de $L(\theta; y_t)$ é o mesmo que encontrar o que maximiza $L(\theta; y_t)$. Portanto, habitualmente, ao invés de encontrar o valor que maximiza $L(\theta; y_t)$, encontra-se o que maximiza $\log(L(\theta; y_t))$- log-verossimilhança [1].

Se as condições de regularidade estiverem satisfeitas, os estimadores de máxima verossimilhança de $\theta_1, \theta_2, ..., \theta_r$ podem ser obtidos através da solução das equações:

$$U_\theta(\theta) = \frac{\partial \log L(\theta; y_t)}{\partial \theta_r} = 0, \tag{2.1.1}$$

i=1,..,r. $U_\theta(\theta)$ é chamado de vetor escore. No entanto, para $\hat{\theta}$ ser de fato o estimador de

[1] Vale acentuar que, neste trabalho, quando se usa $\log(x)$, refere-se ao logaritmo na base natural e (número de Euler).

máxima verossimilhança, é necessário que $U'_{\theta}(\hat{\theta}) < 0$. As propriedades dos estimadores de máxima verossimilhança são:

- Invariância: Se $\hat{\theta}$ é o estimador de máxima verossimilhança de θ, então $g(\hat{\theta})$ é o estimador de máxima verossimilhança de $g(\theta)$;

- Distribuição Assintótica: Sendo $\hat{\theta}$ o estimador de máxima verossimilhança de θ, quando $n \to \infty$, $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N_r(0, I_F(\theta)^{-1})$. Em que, $I_F(\theta)$ é a informação de Fisher esperada que é dada por:

$$I_F(\theta_r) = E\left[\left(\frac{\partial \log L(\theta; y_t)}{\partial \theta_r}\right)^2\right];$$

  Frequentemente, é árduo encontrar a informação de Fisher esperada. Dado isso, utiliza-se a informação de Fisher observada, que é dada por:

$$J(\hat{\theta_r}) = E\left[-\frac{\partial^2 \log L(\hat{\theta}; y_t)}{\partial^2 \theta_r}\right];$$

- Consistência: Para amostras grandes, os estimadores de máxima verossimilhança são não viesados;

- Eficiência: Para amostras grandes, os estimadores de máxima verossimilhança tem a menor variância possível dentre os estimadores não viesados.

2.1.1 Métodos numéricos

Em muitos casos, não é possível encontrar as raízes da equação (2.1.1) de maneira analítica, sendo assim ela deve ser encontrada de maneira numérica. De acordo com José e Furlan (2006), compreende-se como método numérico um algoritmo composto por um número finito de operações envolvendo apenas números (operações aritméticas elementares, cálculo de funções, consulta a uma tabela de valores, consulta a um gráfico, arbitramento de um valor, etc.). Existem diversos métodos numéricos para resolver estas equações, e não é o objetivo deste trabalho discutir todos eles. Entretanto, vale enunciar que os métodos que serão frequentemente utilizados são o de Newtow-Raphson, Escore de Fisher e Algorítimo RS. Cada um desses métodos tem suas vantagens e desvantagens em relação aos outros. Para mais detalhes ver (MCLACHLAN; KRISHNAN, 2007) e (STASINOPOULOS et al., 2015).

2.2 Distribuições

O foco dessa seção é elucidar sobre mistura de distribuições, a distribuição Bernoulli e a distribuição beta. Essas distribuições são de extrema importância para este trabalho, uma vez que a mistura delas nos dará uma outra distribuição cujo o domínio é o intervalo [0,1], distribuição beta inflacionada em zero e um (OSPINA, 2008).

2.2.1 Mistura de Distribuições

Mistura de distribuições são extramente úteis em diversas áreas, já que são extremamente flexíveis e amplamente capazes de descrever a heterogeneidade na distribuição de uma variável. Seja $Y_1, Y_2, ..., Y_n$ uma amostra aleatória de tamanho n e seja Y_t um vetor p-dimensional com função de densidade ou função de probabilidade $f(y_t)$ em $\mathbb{R}^p$. De acordo com McLachlan e Peel (2004), podemos definir a densidade da mistura Y_t como:

$$f(y_t) = \sum_{i=1}^{n} \pi_i f_i(y_t), \tag{2.2.1}$$

em que:

- $0 < \pi_i < 1$ e $\sum_{i=1}^{n} \pi_i = 1$. π_i é chamado de peso ou proporção da mistura;

- $f_i(y_t)$ são chamados componentes de densidades da mistura;

- $f(y_t)$ é uma função convexa. Uma vez que uma combinação linear de densidades de uma variável aleatória pode ou não ser uma função de densidade, conquanto se $f(y_t)$ for uma função convexa, ela garante que a combinação será uma função densidade.

O k-ésimo momento de uma variável aleatória mista é o k-ésimo momento de cada componente (desde que estejam bem definidos) com suas respectivas proporções de mistura, isto é:

$$E(Y^k) = \sum_{i=1}^{n} \pi_i E(Y_i^k). \tag{2.2.2}$$

Por exemplo, considere Y uma variável aleatória sendo uma mistura de uma variável aleatória contínua com uma discreta. Desde que estejam bem definidos, seus momentos serão:

$$E(Y^k) = \pi_D E(Y_D^k) + \pi_C E(Y_C^k), \tag{2.2.3}$$

em que π_D e π_C representam as probabilidades não negativas de selecionar a parte discreta e contínua, respectivamente. Sendo assim, temos que:

- E(Y) = $\pi_D E(Y_D) + \pi_C E(Y_C)$;

- Var(Y) = $\pi_D.Var(Y_D) + \pi_C.Var(Y_C) + \pi_D\pi_C.(E(Y_D) - E(Y_C))^2$.

2.2.2 Distribuição de Bernoulli

A distribuição de Bernoulli é usada em situações que variável só pode assumir valores dicotômicos, como em classificações e previsão de risco, por exemplo.

Dizemos que uma variável aleatória discreta Y segue uma distribuição de Bernoulli de parâmetro p, se ela só assume valores binários, isto é, $\Omega = \{0, 1\}$, que são comumente classificado como "sucesso(1)" e "fracasso(0)" e sua função de probabilidade é dada por:

$$P(Y = y) = \begin{cases} p, & \text{para y=1} \\ 1 - p, & \text{para y=0} \end{cases}, \text{ com } p \in (0, 1).$$

Desta forma, temos que $E(Y) = p$ e $Var(Y) = p(1 - p)$. A notação para indicar que Y segue uma distribuição de Bernoulli é $Y \sim Ber(p)$.

2.2.3 Distribuição beta

A distribuição beta é bastante flexível em problemas no qual a variável resposta está limitada no intervalo (0,1), como taxa, proporções, definição de prioris (GELMAN et al., 2013), etc. Vale relatar que a variável resposta Y não necessariamente precisa estar no intervalo (0,1), caso ela esteja em um intervalo (a,b), a e b $\in \mathbb{R}$ e $b > a$, basta gerar uma nova variável $\tilde{Y}$ através da transformação $\tilde{Y}$= (y-a)/(b-a), que $\tilde{Y}$ estará no intervalo (0,1).

Dizemos que uma variável aleatória Y segue uma distribuição Beta com parâmetros $\alpha > 0$ e $\beta > 0$, se ela estiver no intervalo (0,1) e sua função de densidade é dada por:

$$f(y; \alpha; \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1 - y)^{\beta-1} \quad \forall y \in (0, 1), \tag{2.2.4}$$

em que $\Gamma(j) = \int_0^\infty u^{j-1}e^{-u}du$ é a função gama. A notação para indicar que Y segue uma distribuição Beta é $Y \sim Beta(\alpha, \beta)$.

Neste trabalho será feita uma reparametrização proposta por Ferrari e Cribari-Neto (2004) de modo que a distribuição Beta fique em função de sua média de um parâmetro de dispersão. Portanto, seja $\mu = \alpha/(\alpha+\beta)$ e $\phi = \alpha+\beta$, isto é, $\alpha = \mu\phi$ e $\beta = (1-\mu)\phi$. Desta forma, a densidade de Y reparametrizada será dada por:

$$f(y;\mu;\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \quad \forall y \in (0,1), \tag{2.2.5}$$

com $0 < \mu < 1$ e $\phi > 0$. Nesta parametrização, temos que o k-ésimo momento de Y é dado por

$$\begin{aligned} E(Y^k) &= \int_{-\infty}^{+\infty} y^k f(y;\mu;\phi)dy \\ &= \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} \int_0^1 y^{\mu\phi+k-1}(1-y)^{(1-\mu)\phi-1}dy \\ &= \frac{\Gamma(\phi)\Gamma(\mu\phi+k)\Gamma((1-\mu)\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)\Gamma(\phi+k)} \int_0^1 \frac{\Gamma(\phi+k)y^{\mu\phi+k-1}(1-y)^{(1-\mu)\phi-1}}{\Gamma(\mu\phi+k)\Gamma((1-\mu)\phi)}dy \\ &= \frac{\Gamma(\phi)\Gamma(\mu\phi+k)}{\Gamma(\phi+k)\Gamma(\mu\phi)}. \end{aligned} \tag{2.2.6}$$

Sendo assim, por (2.2.6) temos que:

- $E(Y) = \mu$;
- $Var(Y) = \dfrac{V(\mu)}{(\phi+1)}$,

em que $V(\mu) = \mu(1-\mu)$ é a função de variância. Por consequência, nessa parametrização, o parâmetro μ é a média de Y e, fixado μ, ϕ será o parâmetro de precisão, pois quanto maior ele for, menor será a variância. Para $\mu > 0.5$ e $\phi > 2$, a moda da distribuição beta é dada por:

- Moda(Y)= $\dfrac{\mu\phi-1}{\phi-2}$.

Assim como foi mencionado anteriormente, a distribuição beta é muito flexível, em razão de que com apenas dois parâmetros ela consegue assumir diferentes formas. Por exemplo, quando o parâmetro μ for igual à 0.5, a distribuição será assimétrica independente da escolha do parâmetros ϕ. Quando μ for igual à 0.5 e ϕ for igual à 2, a distribuição será uma Uniforme(0,1). Quando μ for igual à 0.5 e ϕ for igual a 1 a distribuição será a arco seno, etc. A figura abaixo ilustra as diferentes formas da distribuição beta para alguns valores de μ e ϕ.

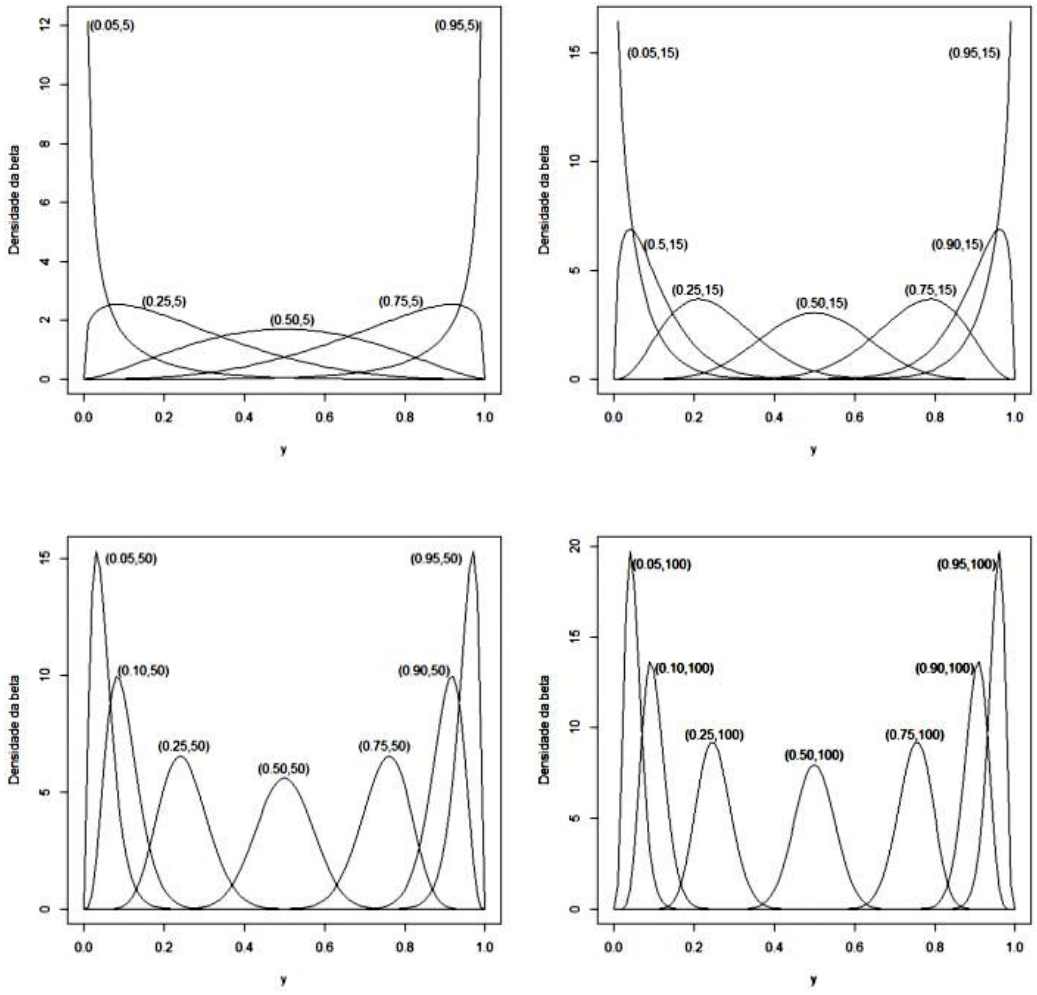


Figura 1: Distribuição beta para diferentes valores de (μ, ϕ).

2.3 Modelos

2.3.1 Modelo de Regressão Linear

Pode-se dizer que o método estatístico mais usado para estabelecer uma relação entre uma variável dependente (resposta) com variáveis explicas é a regressão linear múltipla.

Sendo $Y_1, Y_2, ..., Y_n$ uma amostra aleatória, podemos definir a regressão linear múltipla como:

$$Y_t = \sum_{k=1}^{p} X_{kt}\beta_k + \varepsilon_t, \tag{2.3.1}$$

para t=1,2,...,n, em que:

- X_t são co-variáveis explicativas conhecidas (fixas);

- β_k são parâmetros desconhecidos a serem estimados;

- ε_t são variáveis aleatória i.i.d com distribuição $N(0, \sigma^2)$.

Note que, como Y_t é uma combinação linear de ε_t, que tem distribuição normal, Y_t também segue distribuição normal.

O fato de Y_t seguir uma distribuição normal traz algumas limitações ao modelo de regressão usual. Por exemplo, em alguns casos a variável resposta está limitado à um certo intervalo [a,b] (a e b $\in \mathbb{R}$ e $b > a$), só assume os valores 0 e 1, são dados de contagem, etc. Nesses casos, o modelo de regressão múltipla torna-se inviável, pois o suporte da distribuição Normal é toda a reta do $\mathbb{R}$. Uma alternativa para solucionar esse problema é fazer uma transformação na variável resposta como $\log(Y)$, Y^2, Y^{-1}, Box-Cox, entre outras. Entretanto, ao aplicar essas transformações, perde-se muito na interpretação dos parâmetros.

2.3.2 Modelos Lineares Generalizados

Quando alcançar os pressupostos do modelo de regressão linear (2.3.1) se torna complicado, uma alternativa é propor modelos alternativos. Habitualmente, a alternativa usada é alvitrar um modelo linear generalizado proposto por (NELDER; WEDDERBURN, 1972).

Os modelos lineares generalizados são uma extensão do modelo de regressão linear e, portanto, são mais flexíveis, uma vez que a variável resposta não precisa ser necessariamente Normal, mas sim pertencer a família de distribuição exponencial unidimensional com dispersão.

Seja $Y_1, Y_2, ..., Y_n$ um vetor de variáveis aleatórias com função de densidade (ou de probabilidade) escrita na forma da família exponencial unidimensional com dispersão como:

$$f(y_t; \theta_t, \phi) = \exp\{\phi^{-1}[y_t\theta_t - b(\theta_t)] + c(y_t, \phi)\}, \tag{2.3.2}$$

em que b(.) e c(.) são funções conhecidas, θ é a forma canônica do parâmetro de locação e ϕ é o parâmetro de dispersão. O modelo linear generalizado é caracterizado pelas partes:

- Componente sistemática: $g(\mu_t) = \eta_t$;

- Preditor Linear: Co-variáveis explicativas $x_t = (x_1, x_2, ..., x_n)$ produz um preditor linear linear $\eta_t = x_t^\top \beta$. Em que $\beta_t^\top = (\beta_1, \beta_2, ..., \beta_p)$, p<n, é um vetor de parâmetros

desconhecidos a serem estimados;

- Função de Ligação: A função de ligação $g(.)$ é uma função monótona, diferenciável e que relaciona o preditor linear (η_t) com o valor esperado, $\mu_t = E(Y_t)$, de um dado Y. Alguma das funções de ligação mais usadas estão listadas na tabela abaixo:

Tabela 1: Funções de Ligação.

Função de ligação	g(.)
Identidade	$g(\mu_t) = \mu_t$
Log	$g(\mu_t) = \log(\mu_t)$
Logit	$g(\mu_t) = \log(\frac{\mu_t}{1-\mu_t})$
Probit	$g(\mu_t) = \Phi^{-1}(\mu_t)$
LogComplementar	$g(\mu_t) = \log(1 - \log(1 - \mu_t))$
LogLog	$g(\mu_t) = -\log(-\log(\mu_t))$

Em que $\Phi(.)$ é a função de distribuição de uma normal padrão.

Mesmo sendo bastante flexível, os MLGs, assim como o modelo de regressão linear, possuem restrições. Como, por exemplo, a necessidade da variável resposta pertencer a família exponencial unidimensional com dispersão (2.3.2). Se a variável resposta tiver distribuição Beta (2.2.5) ou distribuição BIZU (3.1.1), por exemplo, elas não se enquadram na classe dos modelos lineares generalizados, visto que elas pertencem a família de distribuição exponencial de dimensão 2 e 4, respectivamente (OSPINA, 2008).

2.3.3 GAMLSS

Quando não há viabilidade de se utilizar uma MLG, repetidamente, o caminho é usar modelos alternativos, como os modelos Aditivos Generalizados para locação, forma e escala(GAMLSS) proposto por Rigby e Stasinopoulos (2005), por exemplo. Os modelos Aditivos Generalizados para locação, forma e escala são modelos semi-paramétricos. Paramétrico, pois a variável resposta requer uma distribuição e semi, porque a modelagem dos parâmetros da distribuição, como funções de variáveis explicativas, pode envolver o uso de funções de suavização não paramétricas.

No GAMLSS, a variável resposta não precisa necessariamente pertencer a família exponencial. A parte sistemática do modelo é expandida para permitir a modelagem não apenas da média (ou localização), mas de outros parâmetros da distribuição de Y como, linear e/ou não linear, paramétrica e/ou aditiva funções não-paramétricas de variáveis

explicativas e/ou efeitos aleatórios. Portanto, o GAMLSS é adequado quando a variável resposta não faz parte da família exponencial ou quando exibe heterogeneidade - por exemplo, onde a escala ou forma da distribuição da variável resposta muda com a(s) variável(s) explanatória(s).

Seja $Y^\top = (Y_1, Y_2, ..., Y_n)$ o vetor da variável resposta de tamanho n com função de densidade $f(y_t|\theta^t)$, em que $\theta^t = (\mu_t, \sigma_t, \nu_t, \tau_t)$. Os dois primeiros parâmetros da distribuição, μ e σ, são geralmente classificados como parâmetros de escala e locação, enquanto os demais parâmetros são classificados como parâmetros de forma. Sendo $g_k(.)$ uma função de ligação monótona conhecida que relaciona os parâmetros da distribuição com as variáveis explicativas, temos que:

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_K} Z_{jk}\gamma_{jk}, \tag{2.3.3}$$

ou seja,

$$\begin{aligned}
g_1(\mu) &= \eta_1 = X_1\beta_1 + \sum_{j=1}^{J_1} Z_{j1}\gamma_{j1}, \\
g_2(\sigma) &= \eta_2 = X_2\beta_2 + \sum_{j=1}^{J_2} Z_{j2}\gamma_{j2}, \\
g_3(\nu) &= \eta_3 = X_3\beta_3 + \sum_{j=1}^{J_3} Z_{j3}\gamma_{j3} \text{ e} \\
g_4(\tau) &= \eta_4 = X_4\beta_4 + \sum_{j=1}^{J_4} Z_{j4}\gamma_{j4},
\end{aligned}$$

em que, $\mu_t, \sigma_t, \nu_t, \tau_t$ são vetores de tamanho n, η_k é o preditor linear também de tamanho n, $\beta^\top = (\beta_{1k}, \beta_{2k}, ..., \beta_{J_k'k})$ é o vetor de parâmetros de tamanho J_k', X_k é uma matriz de delineamento conhecida de ordem n x J_k'(matriz de efeito fixo), Z_{jk} é uma matriz de delineamento conhecida de ordem n x q_{jk}(matriz de efeito aleatório) e γ_{jk} é uma variável aleatória q_{jk} dimensional que assume distribuição $\gamma_{jk} \sim N_{q_{jk}}(0, G_{jk}^{-1})$, onde G_{jk}^{-1} é o inverso (generalizado) de uma matriz simétrica $q_{jk} \times q_{jk}$ $G_{jk} = G_{jk}(\lambda_{jk})$ que pode depender de um vetor de hiper-parâmetros λ_{jk}. Note que em alguns casos não há interesse em modelar alguns parâmetros em função das variáveis explicativas.

3 Regressão Beta inflacionado em zero e um

3.1 Distribuição beta inflacionada em zero e um

Assim como foi citado na introdução, existem casos em que as taxas, proporções não estão em intervalos (0,1), mas sim em intervalos [0,1]. Em outras palavras, as probabilidades de se observar zero e um são positivas. Diante disso, utilizamos a ideia de mistura de distribuição referida na Subseção 2.2.1; misturando a distribuição beta, como componente contínuo, que está definida no intervalo (0,1), e a distribuição de Bernoulli para as ocorrências de zero e um. Portanto, assumimos a função de distribuição desta mistura proposta por (OSPINA, 2008) como:

$$BIZU(y;\alpha,p,\mu,\phi) = \alpha Ber(y;p) + (1-\alpha)F(y;\mu,\phi), \tag{3.1.1}$$

em que $Ber(y;p)$ representa a função de distribuição de uma variável aleatória com distribuição de Bernoulli com parâmetro p , $F(y;\mu,\phi)$ representa a função de distribuição de uma variável com distribuição $Beta(\mu,\phi)$.

O parâmetro α é a proporção da mistura que nos permite combinar as duas distribuições de maneira convexa. A interpretação dele é simples, com probabilidade α a variável Y é selecionada de uma distribuição de Bernoulli e com probabilidade $1-\alpha$ a variável Y é selecionada de distribuição beta.

Dizemos que uma variável aleatória Y segue uma distribuição beta inflacionada em zero e um (BIZU) com parâmetros $0 < \mu, \alpha, p < 1$ e $\phi > 0$, se ela assumir valores no intervalo [0,1] e sua função de densidade gerada pela mistura for:

$$bizu(y;\alpha,p,\mu,\phi) = \begin{cases} \alpha(1-p), & \text{se y=0} \\ \alpha p, & \text{se y=1} \\ (1-\alpha)f(y;\mu,\phi), & \text{se y} \in (0,1), \end{cases} \tag{3.1.2}$$

em que $f(y;\mu,\phi)$ é a função de densidade beta (2.2.5). A notação para indicar que Y segue uma distribuição beta inflacionada em zero e um é $Y \sim BIZU(\alpha,p,\mu,\phi)$.

Desta forma, pela equação (2.2.3) temos que o k-ésimo momento de Y é dado por:

$$E(Y^k) = \alpha p + (1-\alpha)\mu_k, \tag{3.1.3}$$

onde μ_r é o r-ésimo momento ao redor de zero da distribuição $Beta(\mu,\phi)$. Logo,

- $E(Y) = \alpha p + (1-\alpha)\mu$;

- $Var(Y) = \alpha p(1-p) + \dfrac{(1-\alpha)\mu(1-\mu)}{(\phi+1)} + \alpha(1-\alpha)(p-\mu)^2.$

A distribuição BIZU, assim como a distribuição beta, é bastante flexível e assume diferentes formas dependendo dos valores dos parâmetros (α, p, μ, ϕ). Se μ e p forem iguais a 0.5, independente dos valores de α e $\phi > 1$, a distribuição será simétrica. Se μ ou ϕ forem diferente de 1/2, a distribuição será assimétrica. A figura abaixo mostra o gráfico das densidades para diferentes escolhas de μ e ϕ.

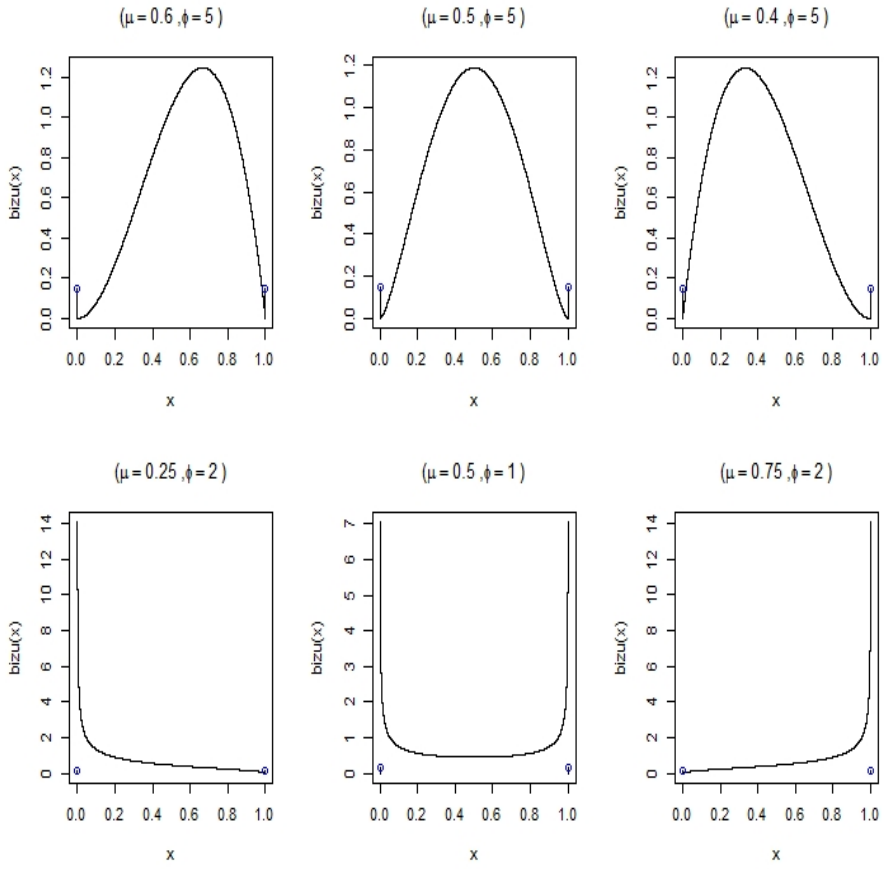


Figura 2: Distribuição BIZU para diferentes valores de μ e ϕ com $\alpha = 0.3$ e $p = 0.5$.

Proposição 3.1 *A distribuição BIZU(3.1.1) pertence a família exponencial de dimensão 4 de posto completo.*

A prova desta proposição está melhor descrita em (OSPINA, 2008). Desta forma, considerando uma amostra aleatória $Y_1, Y_2, ..., Y_n$ de uma distribuição BIZU, seu vetor de estatísticas suficiente completa $\sum_{t=1}^{n} T(y_t) = (T_1, T_2, T_3, T_4)$ é dado por:

$$
\begin{aligned}
T_1 &= \sum_{t=1}^{n} I_{(0,1)}(y_t), \\
T_2 &= \sum_{t=1}^{n} I_{(1)}(y_t), \\
T_3 &= \sum_{t:y_t\in(0,1)}^{n} \log(y_t), \\
T_4 &= \sum_{t:y_t\in(0,1)}^{n} \log(1-y_t),
\end{aligned}
\tag{3.1.4}
$$

em que $I_{(a,b)}(.)$ é a função indicadora.

É agudamente relevante a estimação dos parâmetros da distribuição. Posto isso, temos que a função de máxima verossimilhança de $\theta = (\alpha, p, \mu, \phi)$ é dada por:

$$L(\theta) = \prod_{t=1}^{n} bizu(y_t; \alpha, p, \mu, \phi).$$

Desta maneira, o logaritmo da função de máxima verossimilhança é:

$$l(\theta) = \log(L(\theta)) = l_1(\alpha) + l_2(p) + l_3(\mu, \phi),$$

onde,

$$
\begin{aligned}
&l_1(\alpha) = T_1 \log(\alpha) + (n - T_1)\log(1-\alpha), \\
&l_2(p) = T_2 \log(p) + (T_1 - T_2)\log(1-p) \text{ e} \\
&l_3(\mu,\phi) = (n-T_1)\log\left\{\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}\right\} + T_3(\mu\phi - 1) + T_4((1-\mu)\phi - 1).
\end{aligned}
$$

É importante observar que, l_1 depende apenas de α, l_2 depende apenas de p e l_3 depende apenas de (μ, ϕ). Consequentemente, a máxima verossimilhança pode ser analisada separadamente.

O vetor de escores e o desenvolvimento das equações (2.1.1) estão descritos em (OSPINA, 2008). Assim sendo, os estimadores de máxima verossimilhança de α e p são, respectivamente, $\hat{\alpha} = T_1/n$ e $\hat{p} = T_1/T_2$. Onde, $\hat{\alpha}$ é a proporção de zeros e uns na amostra; é fácil verificar que $\hat{\alpha}$ é um estimador não viesado de variância mínima para α; $\hat{p}$ é a proporção de uns na sub-amostra que contém apenas a massa de zeros e uns. Os estimadores de μ e ϕ não podem ser obtidos de forma analítica, portanto só são obtidos através métodos numéricos.

A matriz de informação de Fisher esperada $I_F(\theta)$ de θ é dada por:

$$I_F(\theta) = \begin{pmatrix} K_{\alpha\alpha} & 0 & 0 & 0 \\ 0 & K_{pp} & 0 & 0 \\ 0 & 0 & K_{\mu\mu} & K_{\mu\phi} \\ 0 & 0 & K_{\phi\mu} & K_{\phi\phi} \end{pmatrix},$$

em que:

$$\begin{aligned} &K_{\alpha\alpha} = 1/\{\alpha(1-\alpha\}, \\ &K_{pp} = \alpha/\{p(1-p\}, \\ &K_{\mu\mu} = (1-\alpha)\phi^2\{\psi^{'}(\mu\phi) + \psi^{'}((1-\mu)\phi)\}, \\ &K_{\mu\phi} = (1-\alpha)\phi\{\psi^{'}(\mu\phi)\mu - \psi^{'}((1-\mu)\phi)(1-\mu)\} \text{ e} \\ &K_{\phi\phi} = (1-\alpha)\{\mu^2\psi^{'}(\mu\phi) + (1-\mu)^2\psi^{'}((1-\mu)\phi) - \psi^{'}(\phi)\}, \end{aligned}$$

onde $\psi(.)$ é a função digama dada por $\dfrac{\Gamma^{'}(.)}{\Gamma(.)}$.

Posto isso, sob as condições de regularidade podemos estabelecer a distribuição assintótica dos parâmetros $\theta = (\mu, \phi, \alpha, \gamma)$. Logo, temos que:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N_4(0, K(\theta)^{-1}),$$

em que $\hat{\theta} = (\hat{\mu}, \hat{\phi}, \hat{\alpha}, \hat{\gamma})$. Podemos também estabelecer intervalos de confiança para os parâmetros. Por exemplo, à um nível de significância (1-α)100%, o intervalo de confiança para μ é dado por,

$$\hat{\mu} \pm z_{1-\frac{\alpha}{2}}(\hat{K}^{\mu\mu})^{1/2},$$

onde $z_{1-\frac{\alpha}{2}}$ é o quantil $(1-\alpha/2)$ da N(0,1) e $\hat{K}^{\mu\mu}$ é o elemento da matriz de informação de Fisher observa referente ao parâmetro μ.

Uma reparametrização proposta por (OSPINA, 2008) da distribuição BIZU será utilizada neste trabalho. Na qual, $\delta_1 = \alpha p$ e $\delta_0 = \alpha(1-p)$. Desta forma, o parâmetro de mistura será dado por $\alpha = \delta_0 + \delta_1$. Portanto, a nova densidade da BIZU será dada por:

$$bizu(y; \delta_0, \delta_1, \mu, \phi) = \begin{cases} \delta_0, & \text{se y=0} \\ \delta_1, & \text{se y=1} \\ (1-\delta_0-\delta_1)f(y;\mu,\phi), & \text{se y} \in (0,1), \end{cases} \tag{3.1.5}$$

em que $f(y;\mu,\phi)$ é a densidade da distribuição beta (2.2.5). Com esta parametrização, a interpretação dos parâmetros é mais simples, visto que $\delta_0 = P(Y=0)$, $\delta_1 = P(Y=1)$ e (μ, ϕ) são os parâmetros da distribuição beta para a probabilidade do intervalo contínuo (0,1).

3.2 Modelo Regressão Beta inflacionado em zero e um

Regularmente quando a variável resposta está limitada em um certo intervalo (0,1), o mecanismo mais utilizado é fazer transformações na mesma para que ela esteja definida em toda a reta. Todavia, ao se fazer isso, se perde muito na interpretação dos parâmetros. Posto isso, alguns autores propuseram modelos em que a variável resposta seguiria uma distribuição beta para a variável resposta, como (FERRARI; CRIBARI-NETO, 2004), por exemplo.

Os modelos de regressão beta se assemelham muito aos MLGs propostos por Nelder e Wedderburn (1972), uma vez que nos permite modelar a média, através de uma função de ligação, usando um preditor linear. Mas, como já foi dito precedentemente, existem casos em que a variável de interesse está no intervalo [0,1), (0,1] ou [0,1]. Com a mesma finalidade de não perder a interpretação dos parâmetros, Ospina (2008) propôs um modelo em que a variável resposta está nestes intervalos, utilizando a distribuição da variável resposta como a beta inflacionada em zero e um (3.1.1). Posteriormente, Pereira (2012) propôs um modelo regressão beta inflacionados truncado em um valor c $\in (0,1)$.

Não obstante, os modelos de regressão beta não podem ser considerado um MLG, uma vez que um dos pressupostos para ser um MLG é que a variável resposta pertença a família exponencial unidimensional com dispersão(2.3.2) e a distribuição beta pertence a família exponencial de dimensão 2. Ora, o mesmo vale para os modelos de regressão de beta inflacionado em zero e um, em razão de que a distribuição beta inflacionada em zero e um pertence a família exponencial de dimensão 4. Outro motivo é que também temos interesse em modelar não só a média como as probabilidades de ocorrência de 0 e 1.

Os modelos de regressão beta inflacionados pertencem à uma classe mais ampla, os modelos Aditivos Generalizados para locação, forma e escala (RIGBY; STASINOPOULOS, 2005), em virtude de que não precisamos restringir a distribuição da varável resposta e que não precisamos apenas utilizar uma função de ligação, como também podemos usar funções de suavização.

Seja $Y_1, Y_2, ..., Y_n$ um vetor de variáveis aleatórias com distribuição BIZU (3.1.5), ou seja, $Y_t \sim$BIZU$(\mu, \phi, \delta_0, \delta_1)$. O modelo de regressão de beta inflacionado em zero e um(RBIZU) é definido por (3.1.5) e pelos componentes sistemáticos:

$$\begin{aligned} g(\mu_t) &= \eta_t = \textstyle\sum_{i=1}^{n} x_{ti}\beta_i \\ H(\delta_{0t}, \delta_{1t}) &= (h_0(\delta_{0t}, \delta_{1t}), h_1(\delta_{0t}, \delta_{1t})) = (\zeta_{0t}, \zeta_{1t}), \end{aligned} \tag{3.2.1}$$

onde $\mu_t = \mathrm{E}(Y_t | Y_t \in (0,1))$, $\delta_0 = \mathrm{P(Y=0)}$, $\delta_1 = \mathrm{P(Y=1)}$ e 1 - δ_0 - $\delta_1 = \mathrm{P}(Y_t \in (0,1))$. $\eta_t = x_t^\top \beta$, $\zeta_{0t} = v_t^\top \rho$ e $\zeta_{1t} = z_t^\top \gamma$ são preditores lineares; $\beta = (\beta_1, \beta_2, ..., \beta_k)^\top$, $\rho = (\rho_1, \rho_2, ..., \rho_{k_0})^\top$ e $\gamma = (\gamma_1, \gamma_2, ..., \gamma_{k_1})^\top$ são os parâmetros da regressão a serem estimados.

E $x_t = (x_{t1}, x_{t2}, ..., x_{tk})^\top$, $v_t = (v_{t1}, v_{t2}, ..., v_{tk_0})^\top$ e $z_t = (z_{t1}, z_{t1}, ..., z_{tk_1})^\top$ são variáveis exógenas conhecidas que representam as dimensões k, k_0 e k_1, respectivamente.

Assume-se que a função de ligação $g : (0,1) \to \mathbb{R}$ é uma função estritamente monótona e duplamente diferenciável. A função H é uma transformação bijetora do conjunto $C = \{(\delta_{0t}, \delta_{1t}) : 0 < \delta_{0t} < 1, 0 < \delta_{1t} < 1 - \delta_{0t}\}$ a $\mathbb{R}^2$ duplamente diferenciável.

A título de exemplo, caso a componente sistemática do modelo RBIZU seja especificada como:

$$\begin{aligned} g(\mu_t) &= \log\left(\frac{\mu_t}{1-\mu_t}\right) \\ H(\delta_{0t}, \delta_{1t}) &= \left(\log\left(\frac{\delta_{0t}}{1-\delta_{0t}-\delta_{1t}}\right), \log\left(\frac{\delta_{1t}}{1-\delta_{0t}-\delta_{1t}}\right)\right). \end{aligned}$$

Consequentemente, temos que

$$\begin{aligned} \delta_{0t} &= P(Y_t = 0) = \frac{\exp(\zeta_{0t})}{1+\exp(\zeta_{0t})+\exp(\zeta_{1t})}, \\ \delta_{1t} &= P(Y_t = 1) = \frac{\exp(\zeta_{1t})}{1+\exp(\zeta_{0t})+\exp(\zeta_{1t}))}, \\ 1-\delta_{0t}-\delta_{1t} &= P(Y_t \in (0,1)) = \frac{1}{1+\exp(\zeta_{0t})+\exp(\zeta_{1t})}. \end{aligned}$$

Neste caso, o modelo RBIZU é chamado de modelo regressão logístico beta inflacionado em zero e um (RLBIZU) (OSPINA, 2008).

3.2.1 Estimação dos parâmetros

A estimação dos parâmetros de regressão do modelo de regressão beta inflacionado em zero e um $\theta = (\rho^\top, \gamma^\top, \beta^\top, \phi)^\top$ é feita através do método da máxima verossimilhança. A função de verossimilhança do modelo é dada por:

$$L(\theta) = \prod_{t=1}^{n} bizu(y_t; \mu, \phi, \delta_0, \delta_1) = L_1(\rho, \gamma) L(\beta, \mu), \tag{3.2.2}$$

em que as porções são dadas por, $L_1(\rho, \gamma) = \prod_{t=1}^{n} \delta_{0t}^{I_{\{0\}}(y_t)} \delta_{1t}^{I_{\{1\}}(y_t)} (1-\delta_{0t}-\delta_{1t})^{1-I_{\{0\}}(y_t)-I_{\{1\}}(y_t)}$ e $L_2(\beta, \phi) = \prod_{t:y_t \in (0,1)} f(y_t; \mu; \phi)$. É importante observar que a inferência por máxima verossimilhança de $(\rho, \gamma)^\top$ e $(\beta, \phi)^\top$ podem ser feitas separadamente; e que $L_1(\rho, \gamma)$ envolve os parâmetros para modelar as ocorrências de 0 e 1, em alternativa $L_0(\beta, \phi)$ envolve os parâmetros para modelar o componente contínuo (0,1).

Dessa forma o log-verossimilhança de θ do modelo RBIZU, é estabelecida como:

$$l(\theta) = \sum_{t=1}^{n} log(bizu(y_t; \mu_t, \phi, \delta_{0t}, \delta_{1t})) = l_1(\rho, \gamma) + l_2(\beta, \gamma), \tag{3.2.3}$$

no qual,

$$\begin{aligned} l_1(\rho, \gamma) &= \sum_{t=1}^{n} l_t(\delta_{0t}, \delta_{1t}), \\ l_2(\beta, \phi) &= \sum_{t:y_t \in (0,1)} l_t(\mu_t, \phi), \end{aligned} \tag{3.2.4}$$

em que,

$$\begin{aligned} l_t(\delta_{0t}, \delta_{1t}) &= I_{\{0\}}(y_t) \log \delta_{0t} + I_{\{1\}}(y_t) \log \delta_{1t} \\ &\quad + (1 - I_{\{0\}}(y_t) - I_{\{1\}}(y_t)) \log(1 - y_t), \\ l_t(\mu_t, \phi) &= \log \Gamma(\phi) - \log \Gamma(\mu_t \phi) - \log((1 - \mu_t)\phi) + (\mu\phi - 1) \log y_t \\ &\quad + \{(1 - \mu_t)\phi - 1\} \log(1 - y_t). \end{aligned} \tag{3.2.5}$$

Em vista disso, precisamos encontrar o vetor de escores dos parâmetros e depois encontrar as soluções das equações $U_\rho(\rho, \gamma) = 0$, $U_\gamma(\rho, \gamma) = 0$, $U_\beta(\beta, \phi) = 0$ e $U_\phi(\beta, \phi) = 0$ (2.1.1) para encontrar os estimadores de máxima verossimilhança. Nada obstante, as soluções destas equações não podem ser obtidas de maneira analítica, em vista disso só poderão ser obtidas de maneira numérica.

Por consequência, satisfeitas as condições de regularidade, podemos definir a distribuição assintótica dos estimadores e fazer testes de hipótese para os estimadores de máxima verossimilhança $\hat{\theta} = (\hat{\rho}_1, ..., \hat{\rho}_{k_0}, \hat{\gamma}_1, ..., \hat{\gamma}_{k_1}, \hat{\beta}_1, ..., \hat{\beta}_k, \phi)$. Portanto, temos que:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N_{k+k_0+k_1+1}(0, I_F(\theta)^{-1}).$$

Como já foi dito, regularmente não é possível obter a informação de Fisher esperada, em vista disso ela é estimada pela informação de Fisher observada- que aqui será denota por $K(\hat{\theta})$. A obtenção dessa matriz está melhor descrita em (OSPINA, 2008). Posto isso, para r=1,...,$k_0 + k_1 + k + 1$, $K(\hat{\theta})^{rr}$ o (r,r)-ésimo elemento da inversa da matriz de informação de Fisher, o intervalo de confiança assintótico de θ para um nível de confiança 100(1-α)% é dado por,

$$\left(\hat{\theta}_r \pm z_{1-\frac{\alpha}{2}} (K(\hat{\theta})^{rr})^{1/2}\right).$$

Usualmente estamos interessados em fazer inferência sobre os parâmetros. Conhecido a distribuição assintótica dos parâmetros, três testes que podemos usar são o teste da razão de log-verossimilhança o teste escore e o teste de Wald. Estes testes estão melhor descrito em (PAULA, 2004) e em (OSPINA, 2008).

A título de exemplo, seja $\rho = (\rho_1^\top, \rho_2^\top)^\top$, $\gamma = (\gamma_1^\top, \gamma_2^\top)^\top$, e $\beta = (\beta_1^\beta, \beta_2^\top)^\top$, onde $\rho_1 = (\rho_1, ..., \rho_{k_{0_1}})^\top$, $\rho_2 = (\rho_{k_{0_1}+1}, ..., \rho_{k_0})^\top$, $\gamma_1 = (\gamma_1, ..., \gamma_{k'})^\top$, $\gamma_2 = (\gamma_{k'+1}, ..., \gamma_{k_1})^\top$, $\beta_1 = (\beta_1, ..., \beta_{k''})^\top$ e $\beta_2 = (\beta_{k''+1}, ..., \beta_k)^\top$. Considere as hipóteses como, $H_0 : \rho_1 = \rho_1^{(0)}; \gamma_1 = \gamma_1^{(0)}; \beta_1 = \beta_1^{(0)}$ contra H_1 : Pelo menos uma igualdade não vale. $\rho_1^{(0)}$, $\gamma_1^{(0)}$ e $\rho_1^{(0)}$ são parâmetros de dimensão k_{0_1}, k_{1_1} e k_1, respectivamente.

A estatística do teste da razão de log-verossimilhança é dada por:

$$\Lambda = 2\{l(\hat{\rho}, \hat{\gamma}, \hat{\beta}, \hat{\phi}) - l(\tilde{\rho}, \tilde{\gamma}, \tilde{\beta}, \tilde{\phi})\},$$

em que $l(.)$ é função log-verossimilhança. e $(\tilde{\rho}, \tilde{\gamma}, \tilde{\beta}, \tilde{\phi})$ é o estimador de máxima verossimilhança de $(\rho, \gamma, \beta, \phi)$ restrito à H_0. Sendo assim, temos que, sob H_0, $\Lambda \xrightarrow{D} \chi^2_{k_{0_1}+k'+k''}$.

3.2.2 Análise de Diagnóstico e Critérios de Informação

Resíduo Quantil Aleatorizado

Para o modelo de regressão beta inflacionado em zero e um consideramos o resíduo quantil aleatorizado

$$r_t^q = \Phi^{-1}(u_t), \qquad t = 1, ..., n, \tag{3.2.6}$$

em que u_t é uma variável aleatória uniforme no intervalo $(a_t,\ b_t]$, sendo $a_t = \lim_{y \uparrow y_t} \text{BIZU}(y; \alpha_t, p_t, \mu_t, \phi)$ e $b_t = \text{BIZU}(y_t; \alpha_t, p_t, \mu_t, \phi)$, respectivamente. Aqui, $\text{BIZU}(y; \alpha_t, p_t, \mu_t, \phi)$ foi definida em 3.1.1 sendo $p_t = \delta_{1t}/\alpha_t$ e $\alpha_t = \delta_{0t} + \delta_{1t}$. O gráfico de r_t^q versus o índice das observações, para que o modelo esteja ajustado, não deve mostrar nenhuma observação fora do padrão detectável. O gráfico de r_t^q versus os valores ajustados, caso apresente uma tendência detectável, pode sugerir que há erro de especificação da função de ligação. Utiliza-se também *QQ-plot* e densidade dos resíduos para verificar a distribuição dos resíduos.

Worm Plot

O *worm plot* (BUUREN; FREDRIKS, 2001) visualiza diferenças entre duas distribuições, condicionadas aos valores de uma covariável. Ele consiste em uma coleção de gráficos *QQ-plot*, com tendência global removida, a fim de evidenciar alterações locais no domínio de uma dada covariável. O eixo vertical do *worm plot* retrata, para cada observação, a diferença entre a sua localização no meio teórico e distribuições empíricas.

Os pontos em cada gráfico formam um *worm*(minhoca); o gráfico resultante tem um formato semelhante a uma curva, responsável pela denominação do gráfico: *worm*(minhoca), o formato dessa curva fornece indícios sobre como os dados diferem da distribuição as-

sumida e sugere possíveis modificações para readequar o ajuste do modelo (JUNIOR, 2017).

A Tabela 2 ilustra as conclusões para as diferentes formas do *worm plot*.

Tabela 2: Interpretação do *worm plot*.

Forma	Momento	Se	Então
Intercepto	Média	a curva está acima da origem,	a média é subestimada
		a curva está abaixo da origem,	a média é superestimada
Inclinação	Variância	a curva tem inclinação positiva	a variância é subestimada
		a curva tem inclinação negativa,	a variância é superestimada
Parábola	Assimetria	a curva tem formato de U,	há excesso de assimetria à esquerda.
		a curva tem formato de U invertido,	há excesso de assimetria à direita.
Curva em S	Curtose	a curva tem forma de S decrescente,	a cauda é excessivamente leve.
		a curva tem forma de S crescente	a cauda é excessivamente pesada.

Detrended Transformed Owen's Plot - DTOP plot.

O *DTOP plot* é baseado na construção de Owen (1995) de faixas de confiança não paramétricas para a função de distribuição verdadeira (dada a função de distribuição empírica da amostra). O método usa uma plotagem de Owen transformada corrigida aplicada aos resíduos quantílicos normalizados (randomizados) do modelo ajustado. O procedimento pode ser aplicado a ambas distribuições, contínuas e discretas. Ele pode ser usado para investigar a adequação da distribuição da variável resposta em uma situação de regressão dentro de intervalos de cada variável explicativa e fornecer orientação para melhorar o modelo.

O gráfico é uma faixa de confiança de probabilidade não paramétrica corrigida para uma função de distribuição. Isto é, se as linhas horizontais estiverem dentro da faixa de confiança, então os resíduos normalizados poderiam ter vindo de uma distribuição Normal e, consequentemente, a distribuição da variável resposta presumida é razoável. A interpretação da forma do gráfico de Owen transformado é análoga à interpretação de *worm plot* dada na Tabela 2. A Tabela 3 abaixo dá a interpretação correspondente para o gráfico de Owen transformado corrigido, enfocando a forma da função de distribuição acumulada empírica corrigida , isto é , $\Phi^{-1}[F_E(\hat{r}_t)]$ - $\hat{r}_t$ versus $\hat{r}_t$ (DJENNAD et al., 2012).

Tabela 3: Interpretação do *DTOP plot*.

Forma do dtecdf	Resíduos	Variável Resposta
Nível: acima da origem	Média muito pequena	Média muito grande
Nível: abaixo da origem	Média muito grande	Média muito pequena
Linha: inclinação positiva	Variância muito pequena	variância muito grande
Linha: inclinação negativa	Variância muito grande	variância muito pequena
Forma de U	Assimetria Negativa	Assimetria muita alta
Forma de U invertido	Assimetria Positiva	Assimetria muito baixa
Forma de S com a esquerda curvada para baixo	Platicúrtica	Calda pesada
Forma de S com a esquerda curvada para cima	Leptocúrtica	Calda leve
Esquerda curvada para baixo (cima)	cauda esquerda curta (longa)	cauda esquerda longa (curta)
Direita curvada para baixo (cima)	cauda direita longa (curta)	cauda direita curta (longa)

Gráfico de Envelope Simulado

Atkinson (1985) propôs a construção, através de simulação de Monte Carlo, de um tipo de banda de confiança através de simulações, a qual denominou envelope. Neste trabalho será utilizado o pacote *hnp* (MORAL; HINDE; DEMÉTRIO, 2017). A técnica utilizada consiste em traçar os valores absolutos ordenados de um modelo diagnóstico versus as estatísticas de ordem esperada de uma distribuição meio-normal, que pode ser aproximado por:

$$\Phi^{-1}\left(\frac{i+n-\frac{1}{8}}{2n+\frac{1}{2}}\right). \tag{3.2.7}$$

A obtenção do envelope simulado para uma plotagem meio normal é simples e consiste em:

1. Ajustar o modelo;
2. Obter as mesmas medidas de diagnóstico de interesse e ordena-las;
3. Simular de 99 (ou mais) variáveis de resposta usando a mesma matriz de modelos, distribuição de erros e estimativas de parâmetros;
4. ajustando o mesmo modelo a cada variável de resposta simulada e extraindo o mesmo diagnósticos do modelo e, novamente, ordenando os valores absolutos;
5. calcular os percentis desejados (por exemplo, 0,025 e 0,975) dos valores de diagnóstico simulados em cada valor da estatística de ordem esperada e utilizá-los para formar o envelope.

Se uma proporção considerável de pontos estiver fora das bandas de confiança, existe evidência de que o modelo ajustado não é adequado.

Critério de Akaike

Para seleção do modelo consegue-se empregar o critério de Akaike (AIC), proposto por Akaike (1974), no qual é definido como:

$$AIC = -2\log(L(\hat{\theta})) + 2k, \tag{3.2.8}$$

em que $L(\hat{\theta})$ é a função de verossimilhança estimada e k é a quantidade de parâmetros. O modelo que apresentar o menor AIC é o que deverá ser selecionado.

Critério de Informação Bayesiano

Alternativamente possuímos o critério de informação Bayesiano (BIC o SBC), proposto por Schwarz et al. (1978), definido por:

$$BIC = -2\log(L(\hat{\theta})) + k * log(n), \tag{3.2.9}$$

em que $L(\hat{\theta})$ é a função de verossimilhança estima,k é a quantidade de parâmetros e n é o tamanho da amostra. O modelo que apresentar o menor BIC é o que deverá ser selecionado.

Pseudo R^2

No modelo regressão linear definido em 2.3.1, um critério de informação comumente utilizado é o R^2- que diz o quanto o modelo explica a variação total da variável resposta. Nessa conjuntura Nagelkerke et al. (1991) propôs uma medida, pseudo R^2, para os modelos em geral, que é dada por:

$$R^2 = \frac{1 - \left\{\frac{L(M_{nulo})}{L(M_{ajustado})}\right\}^{2/n}}{1 - L(M_{nulo})^{2/n}}, \tag{3.2.10}$$

onde $L(M_{nulo})$ e $L(M_{ajustado})$ representam a verossimilhança do modelo apenas com o intercepto e do modelo ajustado, respectivamente.

Se o modelo ajustado predizer perfeitamente, então $R^2 = 1$; no caso oposto, temos que $L(M_{ajustado}) = L(M_{nulo})$, sendo assim, $R^2 = 0$.

4 Metodologia

Com o objetivo de propor um modelo inferencial para verificar quais fatores estão influenciando na proporção de funcionários do sexo feminino nas empresas, será utilizado o modelo de regressão beta inflacionado em zero e um descrito na Subseção 3.2. Fundamentado por esse modelo iremos verificar a significância de cada variáveis disponíveis e propor um modelo bem ajustado através das técnicas de análise de diagnóstico descritas na Subseção 3.2.2. Visto isso, nessa seção será descrito os meios nos quais será permito aplicar a regressão beta inflacionado em zero e um, bem como os materiais e a descrição dele.

4.1 Métodos

Para a realização deste trabalho, todo os gráficos, tabelas, estimações dos parâmetros e diagnóstico do modelo de regressão beta inflacionado em zero e um, serão feitos através do *software* estatístico R (R Core Team, 2018) e (SAS Institute Inc., 2007). Destacando-se neste trabalho o pacote gamlss (STASINOPOULOS; RIGBY, 2018), nele está implementado diversas técnicas de diagnóstico, bem como o modelo de regressão com resposta beta inflacionado em zero e um BEINF()- vale ressaltar que há mais de 50 distribuições disponíveis nele.

Considerando uma variável aleatória com distribuição beta inflacionada em zero e um definida em 3.1.5 e seja um modelo aditivo generalizados para locação, forma e escala definido em 2.3.3, a função BEINF() utiliza a reparametrização no qual, $\mu = \mu$, $\sigma = \frac{1}{\phi+1}$, $\nu = \frac{\delta_0}{\delta_2}$ e $\tau = \frac{\delta_1}{\delta_2}$, em que $\delta_2 = 1 - \delta_0 - \delta_1$. Neste trabalho será considerado as seguintes funções de ligação:

$$
\begin{aligned}
g(\mu_t) &= \log\left(\frac{\mu_t}{1-\mu_t}\right) = x_t^\top \beta, \\
g(\sigma_t) &= \log\left(\frac{\sigma_t}{1-\sigma_t}\right) = w_t^\top \lambda, \\
g(\nu_t) &= \log(\nu_t) = v_t^\top \rho e \\
g(\tau_t) &= \log(\tau_t) = z_t^\top \gamma.
\end{aligned}
$$

Note que ao escolher essas funções de ligação para μ, ν e τ, definimos o modelo como o de regressão logístico beta inflacionado em zero e um descrito em 3.2.2. Note também que na seção 3.2 o modelo foi definido com ϕ, porém em (SIMAS; BARRETO-SOUZA; ROCHA, 2010) podemos ver como a inferência é dada para o caso em que ϕ não está fixo para o modelo de regressão beta que, neste caso, seria dado de maneira análoga para o

RBIZU. É fácil verificar que neste modelo :

$$\mu_t = \frac{\exp(x_t^\top \beta)}{1+\exp(x_t^\top \beta)},$$
$$\delta_{0t} = \frac{\exp(v_t^\top \rho)}{1+\exp(v_t^\top \rho)+\exp(z_t^\top \gamma)},$$
$$\delta_{1t} = \frac{\exp(z_t^\top \gamma)}{1+\exp(v_t^\top \rho)+\exp(z_t^\top \gamma)}e$$
$$\delta_{2t} = \frac{1}{1+\exp(v_t^\top \rho)+\exp(z_t^\top \gamma)}.$$

Antes de enunciar como a interpretação do parâmetros do modelo é dada, vale fixar alguns conceitos. Seja π uma probabilidade de "sucesso" qualquer, definimos a chance ou *odds* como:

$$odds = \frac{\pi}{1-\pi}.$$

Como $\pi \in (0,1)$ temos que o odds é um valor positivo no qual, caso ele seja maior do que 0, a chance de sucesso é maior do que a de fracasso. Por exemplo, considerando $\pi = 0.75$, obtemos que odds = 3 e, portanto, a chance de sucesso é 3 vezes maior que a de fracasso.

Agora considere os seguintes modelos:

$$\log(Y_t) = \log(\theta_0) + \theta_1 \log(X_t). \tag{4.1.1}$$

$$\log(Y_t) = \theta_0 + \theta_1 X_t. \tag{4.1.2}$$

De acordo com Gujarati e Porter (2011), podemos chamar tais modelos de modelo log-log e log-lin. Uma característica vistosa do modelo log-log(4.1.1) é que o coeficiente θ_1 mede a elasticidade de Y em relação a X, ou seja, a variação percentual correspondente a uma dada variação percentual (pequena) de X. Já no modelo log-lin(4.1.2), se multiplicarmos por 100 o parâmetro θ_1, eles nos dará a variação percentual ou taxa de crescimento de Y para uma variação absoluta em X; 100*θ_1 é pode ser dito como a semielasticidade de Y em X.

Definido isso a interpretação dos parâmetros referente a componente contínua μ, e as de inflação δ_0 e δ_1 é dada através da ideia dos modelos log-log e log-lin aplicado a chance de μ, δ_0 e δ_1.

As escolhas das variáveis que serão incluídas ou retiradas do modelo, tanto quanto os teste da análise de diagnóstico, será baseada em um nível significância de 10%, salvo em casos contrários.

4.2 Material

As mulheres trabalham à muito tempo dentro e fora de suas casas. Há uma tradição que acarreta a exclusão das mulheres do mercado de trabalho, visto que, são equivocadamente vinculadas ao trabalho doméstico. Bruschini e Lombardi (2016) apresenta um vasto leque de trabalhos que procura desconstruir a tese de que lugar de mulher é em casa.

Também na literatura encontramos descrita a longa jornada que, tem na figura da mulher a presença da luta por direitos iguais, não somente de salários, mas também por demais direitos e reconhecimento, no fim da década de 1970 e início de 1980, principalmente (SEDLACEK; SANTOS, 1991).

A situação da mulher no mercado de trabalho, mesmo sendo bem mais difícil e com muitos preconceitos, não é pior porque existem tarefas das quais elas possuem as habilidades necessárias e onde os homens não conseguem realizar com a mesma celeridade e destreza. Isso garante a mulher um espaço no mercado de trabalho que dificilmente será ocupado pelo homem (HIRATA, 2002).

Nesse contexto, será empregado um conjunto de dados disponibilizados pelo Instituto de Pesquisa Econômica Aplicada (IPEA). Os dados,se referem a uma amostra aleatória simples de 500 empresas no período de 2016, neles estão contidos as seguintes variáveis:

Tabela 4: Descrição das Variáveis.

Variável	Descrição
prop	Proporção de Mulheres que a empresa possui
empresa	CNPJ - Cadastro Nacional da Pessoa Jurídica
n_contratos	Quantidade de empregados que a empresa possui
share	Market Share - Fração de participação de uma empresa, produto ou serviço no mercado
rotatividade	Taxa de Rotatividade - Taxa média de saída de funcionários, demissões voluntárias e involuntárias,em relação ao número médio de funcionários de uma empresa em determinado período
eng	Variável indicadora informando se a empresa possui ou não engenheiros
renda	Renda Média mensal da empresa
filiais	Número de Filiais que a empresa possui
estudo	Tempo de estudo médio (em anos) dos funcionários da empresa
uf	Unidade da Federação em que a empresa está situada
empr_anos	Idade da empresa em anos
cnae	Classificação Nacional de Atividades Econômicas (dois dígitos)

4.3 Descrição dos Dados

Para compreender melhor a natureza dos dados, de forma que a modelagem proposta seja de fato viável, foi feita uma análise exploratória dos dados, obtendo-se os resultados descritos abaixo.

Na Tabela 5, podemos observar que os valores de máximo e mínimo da variável resposta é 0 e 1, respectivamente. Nessa amostra verificou-se que 8.8% das proporções é 0 e 4.8% é 1. A Tabela mostra também as estimativas pontuais dessa variável.

Tabela 5: Estatísticas descritiva da proporção de mulheres nas empresas.

Min.	Mediana	Média	Máximo	% de zeros	% de uns
0.0000	0.3229	0.3877	1.0000	8.8000	4.8000

A Figura 3 apresenta que a proporção de mulheres está mais concentrada em menos de 0.5. Esses números indicam que a maior parte das empresas possui uma baixa proporção de mulheres no seu quadro de funcionários. As linhas vermelha e azul representam a massa de zeros e uns, respectivamente.

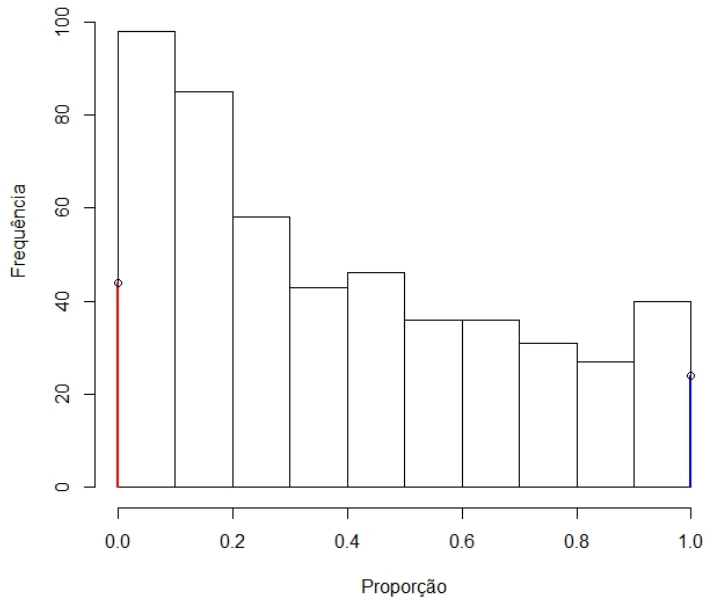


Figura 3: Histograma de frequência para a proporção de mulheres nas empresas.

Seguindo a análise descritiva, segue abaixo a Figura 4 que, mostra a quantidade de

empresas por Unidade da Federação. A figura mostra que há uma concentração de empresas nas regiões sul e sudeste, principalmente no estado de São Paulo. Já a região norte é a que apresenta a menor quantidade de empresas. Por serem as regiões consideradas como as mais desenvolvidas do país, é natural que concentrem um número maior de empresas. Pois são estados que abrigam sedes de importantes multinacionais instaladas no Brasil, convergindo para uma atividade econômica bem mais intensa e, com isso necessitando de mais pessoas para ocuparem vagas no mercado de trabalho.

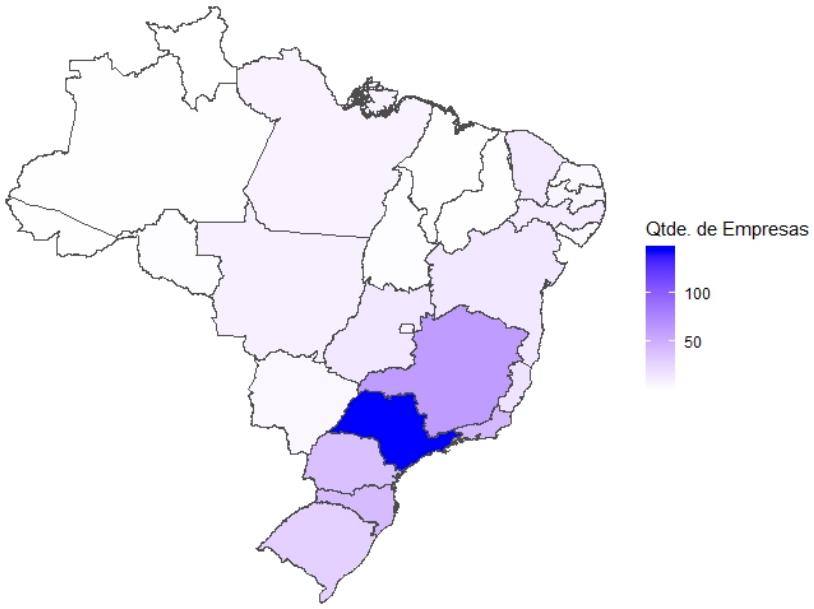


Figura 4: Quantidade de Empresas por UF.

Considerando que há uma disparidade na quantidade de empresas distribuídas pelo país, a Tabela 6 mostra um apanhado geral dos dados relativos ao número de empresas por região. Conforme está explicitado, a região que apresentou o maior número de empresas foi a região sudeste, representando a parcela de 53.8% da amostra, por outro lado a região que apresentou a menor quantidade de empresas foi a região norte, sendo apenas 3.8% da amostra. A região norte não possui como característica o desenvolvimento de atividades relacionadas aos setores secundário e terciário. Sua especialidade é no setor primário, ou seja, atuando como fornecedor de matéria-prima. Logo, é natural que não haja tantas empresas instaladas na região norte, já que o fluxo de pessoas e serviços não é tao intenso, contrastando com a região sudeste que possui o maior número de empresas instaladas em seus estados.

Tabela 6: Frequência de Empresas por Região.

Regiao	Frequência	Porcentagem	E.P. da Porcentagem [2]
Sudeste	269	53.8	2.2318
Sul	110	22.0	1.8544
Nordeste	62	12.4	1.4754
Centro-Oeste	40	8.0	1.2145
Norte	19	3.8	0.8559
Total	500	100.0	

A Tabela 7 mostra o número de empresas por Cnae, assim como o percentual de cada um na amostra. É possível observar que "Comércio; Reparação de Veículos Automotores e Motocicletas" representa 48% das empresas, por outro lado, 0.2% são de "Atividades Imobiliárias". Podemos observar que as classes "Comércio; Reparação de Veículos Automotores e Motocicletas", "Indústrias de Transformação" e "Alojamento e alimentação" juntas contabilizam mais de 80% das classes pertencentes na nossa amostra.

Tabela 7: Frequência de Cnae.

Cnae	Frequência	Porcentagem	E.P. da Porcentagem
COMÉRCIO; REPARAÇÃO DE VEÍCULOS AUTOMO - TORES E MOTOCICLETAS	240	48.0	2.2365
INDÚSTRIAS DE TRANSFORMAÇÃO	114	22.8	1.8781
ALOJAMENTO E ALIMENTAÇÃO	56	11.2	1.4118
TRANSPORTE, ARMAZENAGEM E CORREIO	32	6.4	1.0957
CONSTRUÇÃO	27	5.4	1.0118
ATIVIDADES PROFISSIONAIS, CIENTÍFICAS E TÉCNICAS	14	2.8	0.7385
INFORMAÇÃO E COMUNICAÇÃO	6	1.2	0.4874
AGRICULTURA, PECUÁRIA, PRODUÇÃO FLORESTAL, PESCA E AQUICULTURA	4	0.8	0.3988
ÁGUA, ESGOTO, ATIVIDADES DE GESTÃO DE RESÍDUOS E DESCONTAMINAÇÃO	4	0.8	0.3988
ATIVIDADES FINANCEIRAS, DE SEGUROS E SERVIÇOS RELACIONADOS	2	0.4	0.2826
ATIVIDADES IMOBILIÁRIAS	1	0.2	0.2
Total	500	100.00	

Observando a Tabela 8 percebe-se que, 435 empresas não possuem filiais, o que equivale à 87% das empresas em estudo. Apenas 3.8% das empresas da nossa amostra possui mais de uma filial. É importante analisar também que na nossa amostra existem 2 empresas com 17 filiais, o que, nesse caso, foi um resultado incomum.

[2]Aqui E.P. significa Erro Padrão.

Tabela 8: Frequência de Filiais das Empresas.

Filiais	Frequência	Porcentagem	E.P. da Porcentagem
0	435	87.0	1.505
1	46	9.2	1.294
2	7	1.4	0.526
3	5	1.0	0.445
4	1	0.2	0.200
5	1	0.2	0.200
8	1	0.2	0.200
11	1	0.2	0.200
17	2	0.4	0.282
20	1	0.2	0.200
Total	500	100.00	

No histograma da Figura 5 podemos observar que os valores mais comuns de tempo de estudo médio estão no intervalo de 10-11 e podemos notar uma leve assimetria com concentração à direita. Com isso, conclui-se que é necessário um tempo maior de estudo e qualificação, resultando em maior qualidade no exercício das tarefas por parte dos funcionários. O fato de apresentar uma menor quantidade de empresas, cujo os colaboradores tenham pouco tempo de estudo pode ser um indicativo de que, sem mão-de-obra qualificada, as empresas não conseguem manter a competitividade no mercado e acabam fechando as portas.

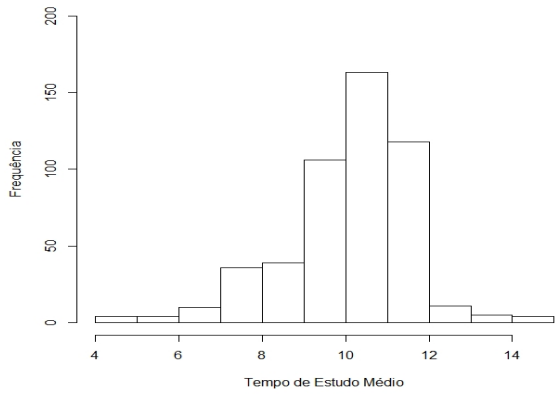
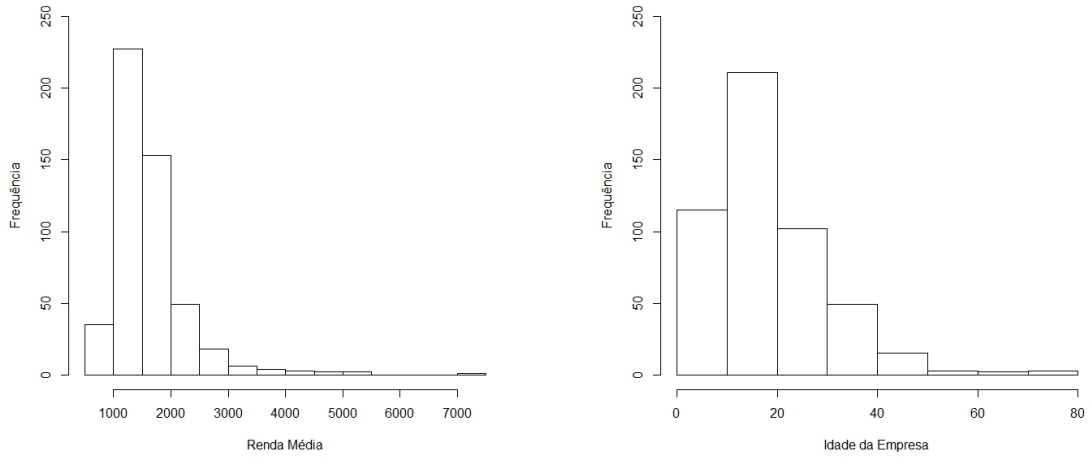


Figura 5: Histograma do Tempo de Estudo Médio da Empresa.

No histograma da Figura 6 (a), constatamos que os valores mais comuns estão no intervalo 1000 a 1500 e uma assimetria com concentração à esquerda, que mostra que valores elevados de renda média são menos comuns, vale ressaltar o valor atípico observado

maior que 7000. Empresas com renda média alta são menos frequentes, o que pode ser explicado por uma série de fatores político-econômicos. Um outro motivo possível pode explicar a renda média da empresa é quantidade de funcionários que ela possui.

No histograma da Figura 6 (b), notamos que os valores mais comuns de idades da empresa estão no intervalo de 10 a 20 anos, podemos notar uma assimetria com concentração à esquerda, isto é, que há uma maior concentração de empresas com idades menores. Essa informação mostra que há uma possível crescente no número de empresas que são criadas, mas também remete à ideia de que empresas com mais tempo de mercado possuem maiores dificuldades para se manterem. Os incentivos dados pelo governo nos últimos 20 ano podem ser fatores que ajudam a explicar o maior número de empresas com essa idade.



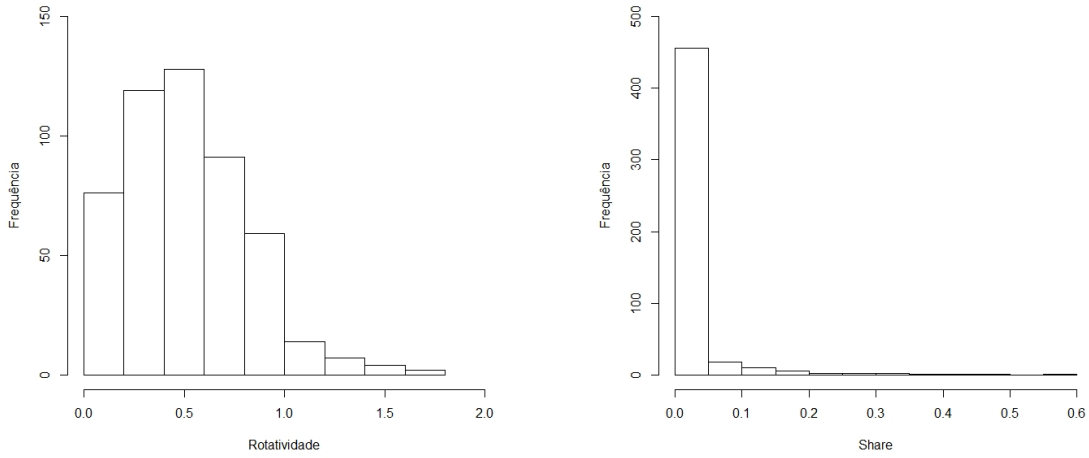
(a) (b)

Figura 6: Histograma da Renda Media da Empresa (a) Histograma da Idade da Empresa (b).

No histograma da Figura 7 (a), observamos uma assimetria com concentração à esquerda, isto é, altos valores observados de rotatividades das empresas são mais raros. Isso é explicado, dentre vários fatores, pelos altos custos existentes para substituição de funcionários, pois é um movimento que gera encargos trabalhistas e impacta no resultado líquido das empresas. Além disso, funcionários com mais tempo de casa, geralmente possuem grande conhecimento acerca das atividades desenvolvidas pela empresa e sobre o dia-a-dia da mesma.

No histograma da Figura 7 (b), uma forte assimetria com concentração à esquerda, isto é, valores elevados de Grau de participação da empresa no mercado em número de vendas de um produto são raros na nossa amostra. A concorrência no mercado dificulta o

domínio de uma única empresa em um determinado setor. Logo, são poucas empresas que conseguem se sobressair em número de vendas em um determinado setor já que o leque de opções disponíveis aos consumidores é cada vez mais amplo. Além disso, há empresas que procuram diversificar suas atividades e atuar em mais de uma área no mercado, preferindo ter pequenas participações porém em vários tipos de atividades.



(a) (b)

Figura 7: Histograma da Rotatividade da Empresa (a) Histograma do Share da Empresa (b).

As Tabelas 9 e 10 apresentam os resultados referentes à correlação entre as variáveis do estudo. Indicam a intensidade na qual essas variáveis estão correlacionadas entre si, assim como o respectivo p-valor da correlação entre as mesmas. Note que, como o tamanho da amostra é relativamente "grande", o teste rejeita a facilmente a hipótese de que não há correlação linear entre as variáveis.

Notamos uma correlação linear moderada entre as variáveis Nº de Contratos e Market Share e Nº de Contratos e Quantidade de Filiais. Esse resultado é de fato esperado, dado que quanto maior o nº de funcionários da empresa maior será a sua representatividade no mercado e também quanto mais filiais a empresa possuir mais funcionários ela terá. As demais variáveis não mostraram ter uma relação linear moderada ou forte entre si.

Tabela 9: Correlação entre as variáveis.

Variáveis	n_contratos	idade	share	rotatividade	estudo	renda	filiais	prop
n_contratos	1.00	0.11	0.61	0.06	-0.01	0.24	0.38	-0.12
idade	-	1.00	0.04	-0.15	-0.08	0.18	0.08	-0.00
share	-	-	1.00	0.03	-0.01	0.17	0.23	-0.12
rotatividade	-	-	-	1.00	-0.06	-0.07	0.08	0.07
estudo	-	-	-	-	1.00	0.26	0.09	0.24
renda	-	-	-	-	-	1.00	0.11	-0.15
filiais	-	-	-	-	-	-	1.00	0.05
prop	-	-	-	-	-	-	-	1.00

Tabela 10: P-valor da Correlação entre as variáveis.

Variáveis	n_contratos	idade	share	rotatividade	estudo	renda	filiais	prop
n_contratos		0.0177	< 0.0001	0.1856	0.8396	< 0.0001	< 0.0001	0.0070
idade	-		0.4180	0.0006	0.0743	< 0.0001	0.0780	0.9285
share	-	-		0.4881	0.8837	0.0001	< 0.0001	0.0067
rotatividade	-	-	-		0.2012	0.1028	0.0908	0.1031
estudo	-	-	-	-		< 0.0001	0.0399	< 0.0001
renda	-	-	-	-	-		0.0103	0.0009
filiais	-	-	-	-	-	-		0.2327
prop	-	-	-	-	-	-	-	

Visando uma melhor compreensão dos dados, as medidas básicas referentes às variáveis da amostra (mínimo, máximo, média, desvio padrão, quantis, além do intervalo de confiança para a média e os respectivos quantis) serão apresentadas nas Tabelas 11 e 12.

Tabela 11: Análise descritiva das variáveis.

Variável	Media	EP	IC 95% p/ Média		C.V. [3]
Nº de Contratos	29.375	2.574	24.318	34.432	0.0876
Idade	18.866	0.489	17.904	19.828	0.0259
Share	0.019	0.002	0.014	0.024	0.1287
Rotatividade	0.515	0.014	0.488	0.542	0.0270
Estudo	10.054	0.066	9.924	10.185	0.0065
Renda	1628.063	29.892	1569.334	1686.794	0.0183
Prop	0.388	0.0135	0.361	0.414	0.0348

[3]Aqui C.V. significa Coeficiente de Variação.

Tabela 12: Análise descritiva das variáveis.

Variável	Mediana	IC 95%	Min	Max
Nº de Contratos	12	(10.905;13.095)	5.000	528.573
Idade	16.567	(15.615;17.518)	5.575	76.000
Share	0.00286	(0.002,0.003)	0.000345	0.550
Rotatividade	0.473	(0.446,0.500)	0	1.692
Estudo	10.367	(10.210,10.525)	4.271	14.956
Renda	1458.83	(1403.017,1514.643)	697.582	7098.566
Prop	0.322	(0.274,0.371)	0	1.000

Abaixo, a Figura 8 que mostra a proporção de mulheres nas empresas por UF. A Figura mostra que na região Norte a média proporção de mulheres nas empresas é superior as demais regiões. Podemos observar que, no país, os estados que apresentam uma proporção média de mulheres nas empresas é o estado de Amazonas, Roraima e Rio Grande do Norte. Observamos também que o estado do Piauí e de Sergipe apresentaram uma proporção média pequena. Podemos observar também que as regiões Centro-Oeste e Sul têm proporções semelhantes distribuídas em seus estados. Vale ressaltar que estas proporções são influenciadas pelas quantidades de empresas por UF pertencentes na amostra descrito na Figura 4.

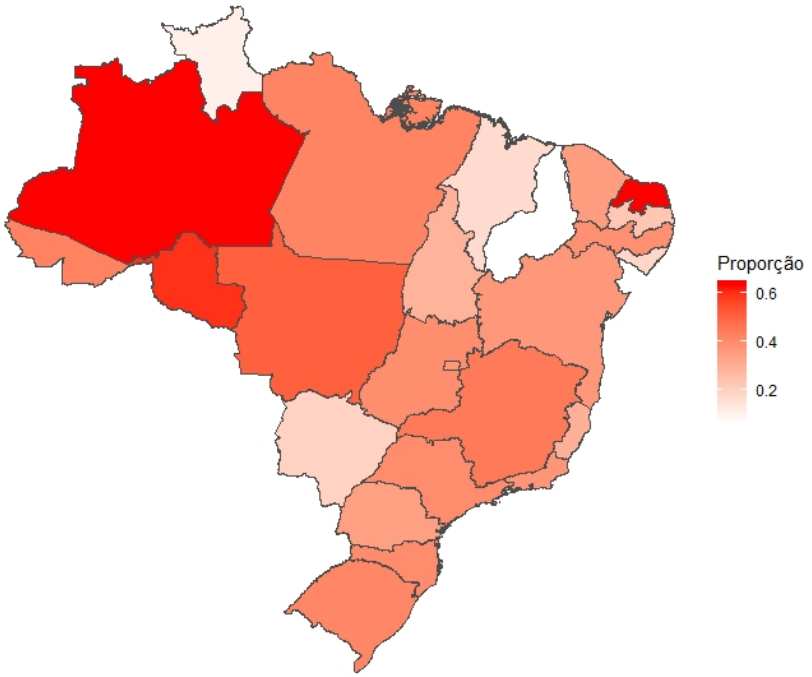
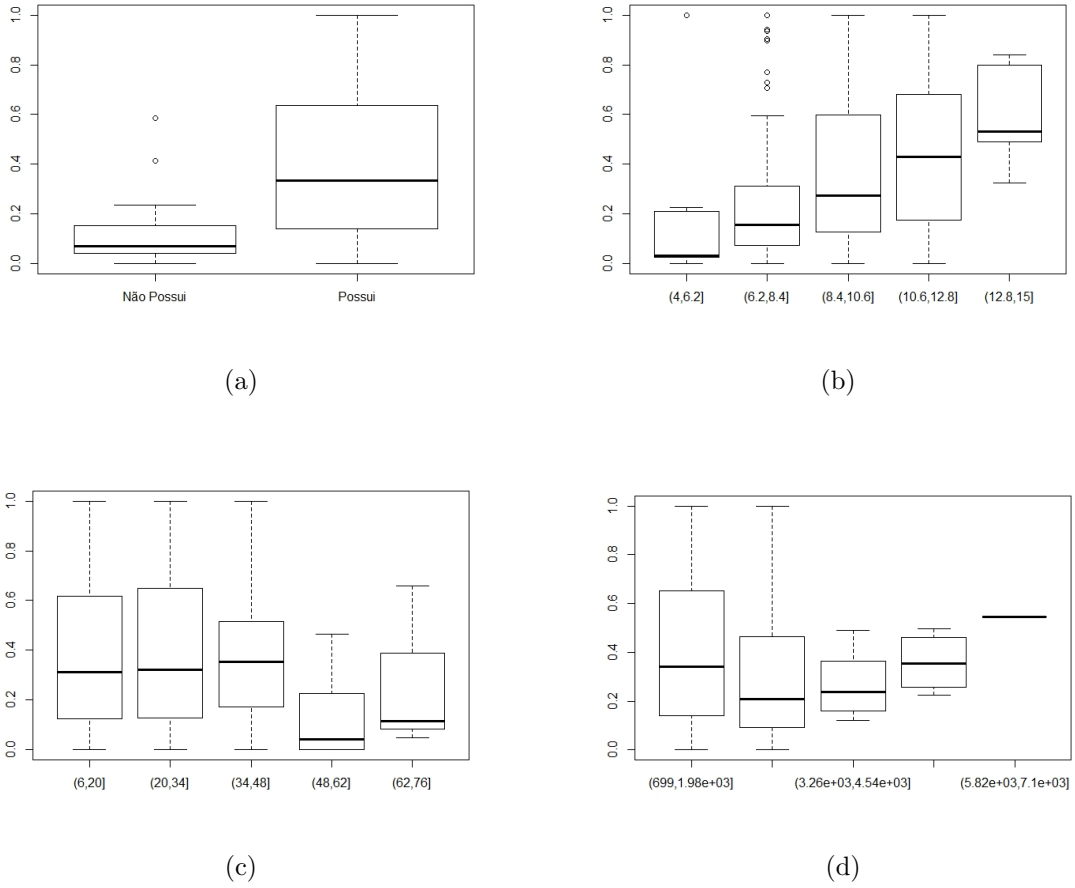


Figura 8: Média da proporção de mulheres nas empresas por UF.

Por meio da Figura 9 podemos observar como a proporção de mulheres na empresas se dá em função das variáveis. Podemos observar que a proporão é maior para as empresas que possuem engenheiros. Observa-se também que a proporção de mulheres nas empresas cresce conforme o tempo de estudo média aumenta. Observamos também que não há crescimento ou decrescimento notável na proporção dado a idade da empresa.



(a) (b)

(c) (d)

Figura 9: Boxplot da proporção por: Possui engenheiros (a)Tempo estudo médio(b) Faixa de idade (c) Faixa de renda média (d).

5 Resultados e Discussões

Devido a natureza da nossa variável de interesse, vimos que o modelo de regressão beta inflaciona em zero e um pode ser uma alternativa que se adéque bem ao problema. Neste capítulo será mostrado todo processo de seleção do modelo, ajuste e interpretação do mesmo.

5.1 Seleção do Modelo

Vimos através da análise descritiva dos dados como a variável resposta (proporção de mulheres nas empresas) e as explicativas estão arranjadas. Com base nesses resultados, algumas considerações são importantes antes de iniciar a modelagem.

Primeiramente, sabe-se que ao trabalhar com uma variável categórica, com z categorias, podemos expressa-la no modelo com $z-1$ categorias, sendo a categoria faltante a de referência. Entretanto, caso o número de categorias seja muito grande, é árduo analisar um modelo com todas elas. Vimos também que variáveis UF e CNAE nessa amostra apresentam uma frequência alta e algumas categorias e baixas nas outras, portanto essas variáveis serão agrupadas.

No tocante a UF, será utilizado a região a qual a empresa está situada. Vimos pela Tabela 6 que a maioria das empresas pertencentes a amostra estão situadas na região sudeste - o que já era de se esperar devido ao cenário econômico-demográfico brasileiro. Dado isso, a região sudeste será escolhida como categoria de referência neste estudo. Já com relação a CNAE, notamos pela Tabela 7 que as categorias "Comércio; Reparação de Veículos Automotores e Motocicletas", "Industrias de Transformação" e "Alojamento e Alimentação" representam juntas 82% da amostra, enquanto as demais apenas 18%. Por consequência, essas demais categorias foram classificadas como outras e também como a categoria de referência.

Nota-se pela análise descritiva dos dados que as variáveis *renda* e *share* apresentam valores muitos altos e baixos, respectivamente. Por isso, será usado *log(renda)* e *log(share)* ao invés das variáveis originais.

Vimos na Subseção 2.1, que os estimadores de máxima verossimilhança são assintoticamente consistentes. Porém, em geral, estes estimadores são viesados para o real valor do parâmetro. Nesse trabalho, estamos utilizando um amostra de tamanho 500, o que é consideravelmente grande, sendo assim o viés não será um problema sério, portanto aqui não será feita nenhuma correção deste tipo. Todavia, existem soluções para este impasse como, por exemplo, *bootstrap* (OSPINA, 2008).

Por fim, neste trabalho não temos interesse em fazer inferência sobre o parâmetro σ. Aqui ele só será utilizado para "controlar" a dispersão. Portanto à ele só será atribuído a variável *offset* dos dados (nº de contratos) para modelá-lo.

Dado essas considerações, foi-se criado um modelo mais completo possível, isto é, para cada parâmetro foi atribuído todas as variáveis disponíveis, ou seja,

$$\mathbf{log}\left(\frac{\boldsymbol{\mu}}{\mathbf{1-\boldsymbol{\mu}}}\right) = \beta_0 + \beta_1 idade + \beta_2 \log(share) + \beta_3 rotatividade + \beta_4 eng + \beta_{5.1} CentroOeste +$$
$$\beta_{5.2} Nordeste + \beta_{5.3} Norte + \beta_{5.4} Sul + \beta_6 estudo + \beta_7 \log(renda) + \beta_{8.1} I.T. + \beta_{8.2} A.A. +$$
$$\beta_{8.3} C; R.A.M. + \beta_9 filiais + \beta_{10} n\_contratos$$
$$\mathbf{log}\left(\frac{\boldsymbol{\sigma}}{\mathbf{1-\boldsymbol{\sigma}}}\right) = \lambda_0 + \lambda_1 n\_contratos$$
$$\mathbf{log}(\boldsymbol{\nu}) = \rho_0 + \rho_1 idade + \rho_2 rotatividade + \rho_3 eng + \rho_{4.1} CentroOeste + \rho_{4.2} Nordeste +$$
$$\rho_{4.3} Norte + \beta_{4.4} Sul + \rho_5 estudo + \rho_{6.1} I.T. + \rho_{6.2} A.A. + \rho_{6.3} C; R.A.M. + \rho_7 filiais +$$
$$\rho_8 n\_contratos + \rho_9 log(renda) + \rho_{10} log(share)$$
$$\mathbf{log}(\boldsymbol{\tau}) = \gamma_0 + \gamma_1 idade + \gamma_2 rotatividade + \gamma_3 eng + \gamma_{4.1} CentroOeste + \gamma_{4.2} Nordeste +$$
$$\gamma_{4.3} Norte + \gamma_{4.4} Sul + \gamma_5 estudo + \gamma_{6.1} I.T. + \gamma_{6.2} A.A. + \gamma_{6.3} C; R.A.M. + \gamma_7 filiais +$$
$$\gamma_8 n\_contratos + \gamma_9 log(renda) + \gamma_{10} log(share).$$

[4]Definido o modelo, foi obtido as estimativas de cada parâmetro e seus respectivos p-valores - como mostra a Tabela 13. Nela podemos ver quais variáveis devem ser retiradas ou não do modelo.

Em relação ao parâmetro μ, foi-se retirado as variáveis uma a uma e checando quais delas eram significativas ou não. Fazendo isso, notou-se que as variáveis idade e rotatividade não mostraram serem significativas. Pode-se ver também que a categoria Nordeste da variável região mostrou ser significativa, diferente das demais categorias, independente disso, essa variável continuou no modelo. As demais variáveis se mostraram significativas e, portanto, prosseguiram no modelo.

De maneira análoga, foi feito o mesmo com os demais parâmetros. Com referência ao parâmetro ν, notamos que as variáveis *eng*, *regiao*, *filiais*, *log(renda)* e *log(share)* não demonstraram ser significativas, logo foram retiradas do modelo. As demais variáveis demonstraram ser significativas e, sendo assim continuaram no modelo. Quanto ao parâmetro τ, observamos que as variáveis *idade*, *log(share)*, *rotatividade*, *eng*, *região*, *estudo* e *cnae* não apontaram serem significativas, por consequência foram retiradas do modelo. As variáveis *filiais*, *n_contratos* e *log(renda)* apontaram serem significativas, por isso seguiram adiante no modelo.

[4]Aqui I.T., A.A., C.;R.A.M. referem-se à Indústria de Transformação, Alojamento e Alimentação e Comércio; Reparação de Veículos Automotores e Motocicletas, respectivamente

Tabela 13: Modelo com todas as variáveis.

Modelo	Variável	Parâmetro	Estimativa	Erro Padrão	Pr(> |t|)
Modelo Saturado			$\boldsymbol{\mu}$		
	Intercepto	β_0	0.311	1.310	0.812
	idade	β_1	$4.79*10^{-3}$	$4.27*10^{-3}$	0.262
	log(share)	β_2	-0.136	0.0428	**0.002**
	rotatividade	β_3	0.091	0.155	0.558
	eng(Possui)	β_4	0.742	0.261	**0.004**
	regiao(Centro-Oeste)	$\beta_{5.1}$	-0.075	0.181	0.677
	regiao(Nordeste)	$\beta_{5.2}$	-0.484	0.158	**0.002**
	regiao(Norte)	$\beta_{5.3}$	-0.268	0.245	0.275
	regiao(Sul)	$\beta_{5.4}$	-0.022	0.114	0.844
	estudo	β_6	0.281	0.036	**< 0.001**
	log(renda)	β_7	-0.749	0.172	**< 0.001**
	cnae(I.T.)	$\beta_{8.1}$	0.547	0.15	**< 0.001**
	cnae(A.A.)	$\beta_{8.2}$	0.913	0.183	**< 0.001**
	cnae(C;R.A.M.)	$\beta_{8.3}$	0.293	0.138	**0.034**
	filiais	β_9	0.02	0.023	0.381
	n_contratos	β_{10}	$1.21*10^{-3}$	$8.06*10^{-4}$	0.135
			$\boldsymbol{\sigma}$		
	Intercepto	λ_0	-0.065	0.052	0.214
	n_contratos	λ_1	$-2.73*10^{-3}$	$8.25*10^{-4}$	**0.001**
			$\boldsymbol{\nu}$		
	Intercepto	ρ_0	5.886	5.331	0.27
	idade	ρ_1	-0.026	0.018	0.145
	rotatividade	ρ_2	-1.712	0.642	**0.007**
	eng(Possui)	ρ_3	-0.471	1.238	0.703
	regiao(Centro-Oeste)	$\rho_{4.1}$	0.512	0.551	0.353
	regiao(Nordeste)	$\rho_{4.2}$	0.048	0.500	0.922
	regiao(Norte)	$\rho_{4.3}$	-0.029	1.112	0.978
	regiao(Sul)	$\rho_{4.4}$	-0.803	0.558	0.151
	estudo	ρ_5	-0.354	0.111	**0.001**
	cnae(I.T.)	$\rho_{6.1}$	-1.04	0.529	**0.049**
	cnae(A.A.)	$\rho_{6.2}$	-2.874	1.118	**0.01**
	cnae(C;R.A.M.)	$\rho_{6.3}$	-1.233	0.483	**0.011**
	filiais	ρ_7	0.115	0.425	0.786
	n_contratos	ρ_8	-0.085	0.032	**0.008**
	log(renda)	ρ_9	-0.076	0.643	0.905
	log(share)	ρ_{10}	0.027	0.182	0.878
			$\boldsymbol{\tau}$		
	Intercepto	γ_0	4.695	$1.70*10^3$	0.998
	idade	γ_1	-0.003	0.026	0.889
	rotatividade	γ_2	-0.332	0.769	0.666
	eng(Possui)	γ_3	8.382	379.930	0.982
	regiao(Centro-Oeste)	$\gamma_{4.1}$	-1.038	1.095	0.343
	regiao(Nordeste)	$\gamma_{4.2}$	-0.9	0.781	0.25
	regiao(Norte)	$\gamma_{4.3}$	0.863	0.879	0.326
	regiao(Sul)	$\gamma_{4.4}$	0.595	0.582	0.307
	estudo	γ_5	-0.145	0.168	0.387
	cnae(I.T.)	$\gamma_{6.1}$	1.526	1.179	0.196
	cnae(A.A.)	$\gamma_{6.2}$	0.495	1.226	0.686
	cnae(C;R.A.M.)	$\gamma_{6.3}$	0.418	1.151	0.716
	filiais	γ_7	0.611	0.236	**0.01**
	n_contratos	γ_8	-0.086	0.055	0.117
	log(renda)	γ_9	-2.753	1.024	**0.007**
	log(share)	γ_{10}	-0.46	0.331	0.165

Retiradas as variáveis que não mostraram ser significativas no modelo com todas as variáveis, foi verificado como as demais reagem na ausência delas. Com esse objetivo, o novo modelo estabelecido (Modelo 1) é descrito como:

$$\mathbf{log}\left(\frac{\boldsymbol{\mu}}{\mathbf{1-\boldsymbol{\mu}}}\right) = \beta_0 + \beta_2 \log(share) + \beta_4 eng + \beta_{5.1} CentroOeste + \beta_{5.2} Nordeste + \beta_{5.3} Norte +$$
$$\beta_{5.4} Sul + \beta_6 estudo + \beta_7 \log(renda) + \beta_{8.1} I.T. + \beta_{8.2} A.A. + \beta_{8.3} C; R.A.M. + \beta_{10} n\_contratos$$
$$\mathbf{log}\left(\frac{\boldsymbol{\sigma}}{\mathbf{1-\boldsymbol{\sigma}}}\right) = \lambda_0 + \lambda_1 n\_contratos$$
$$\mathbf{log}(\boldsymbol{\nu}) = \rho_0 + \rho_1 idade + \rho_2 rotatividade + \rho_5 estudo + \rho_{6.1} I.T. + \rho_{6.2} A.A. + \rho_{6.3} C; R.A.M. +$$
$$\rho_8 n\_contratos$$
$$\mathbf{log}(\boldsymbol{\tau}) = \gamma_0 + \gamma_7 filiais + \gamma_8 n\_contratos + \gamma_9 log(renda).$$

Podemos notar, pela Tabela 14 Modelo 1, que todas as variáveis foram significativas em todos os parâmetros, com exceção das categorias Centro-Oeste, Norte e Sul referentes à variável região do parâmetro μ. Porém, como a categoria Nordeste se mostrou significativa, não foi optado retirar a variável região do modelo. Como alternativa para contornar essa adversidade, foi optado por agrupar essas variáveis não significativas em apenas uma categoria (Outras) - mantendo a categoria Sudeste como a de referência. Portanto, temos que o próximo modelo (Modelo 2) é dado por:

$$\mathbf{log}\left(\frac{\boldsymbol{\mu}}{\mathbf{1-\boldsymbol{\mu}}}\right) = \beta_0 + \beta_2 \log(share) + \beta_4 eng + \beta_{5.2} Nordeste + \beta_{5.3} Outras + \beta_6 estudo +$$
$$\beta_7 \log(renda) + \beta_{8.1} I.T. + \beta_{8.2} A.A. + \beta_{8.3} C; R.A.M. + \beta_{10} n\_contratos$$
$$\mathbf{log}\left(\frac{\boldsymbol{\sigma}}{\mathbf{1-\boldsymbol{\sigma}}}\right) = \lambda_0 + \lambda_1 n\_contratos$$
$$\mathbf{log}(\boldsymbol{\nu}) = \rho_0 + \rho_1 idade + \rho_2 rotatividade + \rho_5 estudo + \rho_{6.1} I.T. + \rho_{6.2} A.A. + \rho_{6.3} C; R.A.M. +$$
$$\rho_8 n\_contratos$$
$$\mathbf{log}(\boldsymbol{\tau}) = \gamma_0 + \gamma_7 filiais + \gamma_8 n\_contratos + \gamma_9 log(renda).$$

Observamos pela Tabela 14 Modelo 2 que, mesmo após recategorizar a variável região, a categoria Outras não apresentou ser significativa, porém Nordeste ainda é. Dado isso, a variável região será mantida no modelo, mesmo a categoria Outras não sendo significativa. No mais, todas as demais apresentaram ser significativas para todos os parâmetros. Sendo assim, o Modelo 2 será o escolhido.

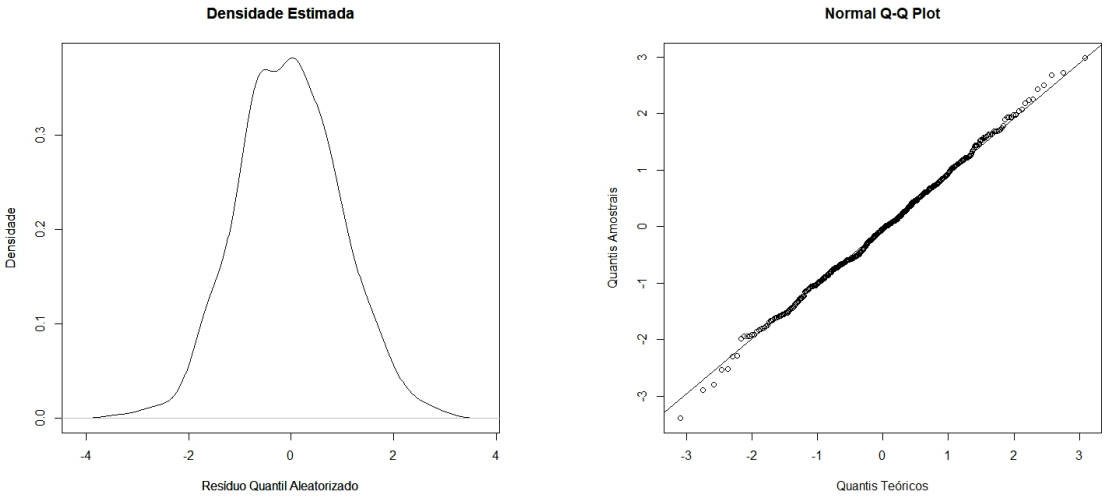
Tabela 14: Seleção dos Modelos.

Modelo	Variável	Parâmetro	Estimativa	Erro Padrão	Pr(> |t|)
Modelo 1			$\boldsymbol{\mu}$		
	Intercepto	β_0	0.234	1.262	0.853
	log(share)	β_2	-0.127	0.041	**0.002**
	eng(Possui)	β_4	0.841	0.247	**< 0.001**
	regiao(Centro-Oeste)	$\beta_{5.1}$	-0.04	0.176	0.818
	regiao(Nordeste)	$\beta_{5.2}$	-0.494	0.155	**0.001**
	regiao(Norte)	$\beta_{5.3}$	-0.258	0.246	0.294
	regiao(Sul)	$\beta_{5.4}$	0.014	0.111	0.902
	estudo	β_6	0.276	0.035	**< 0.001**
	log(renda)	β_7	-0.723	0.168	**< 0.001**
	cnae(I.T.)	$\beta_{8.1}$	0.544	0.147	**< 0.001**
	cnae(A.A)	$\beta_{8.2}$	0.939	0.182	**< 0.001**
	cnae(C;R.A.M.)	$\beta_{8.3}$	0.334	0.133	**0.012**
	n_contratos	β_{10}	$1.54*10^{-3}$	$7.29*10^{-4}$	**0.034**
			$\boldsymbol{\sigma}$		
	Intercepto	λ_0	-0.065	0.052	0.293
	n_contratos	λ_1	$-2.96*10^{-3}$	$8.11*10^{-4}$	**0.001**
			$\boldsymbol{\nu}$		
	Intercepto	ρ_0	4.328	1.298	**0.001**
	idade	ρ_1	-0.029	0.017	**0.098**
	rotatividade	ρ_2	-1.664	0.604	**0.006**
	estudo	ρ_5	-0.320	0.105	**0.002**
	cnae(I.T.)	$\rho_{6.1}$	-1.088	0.497	**0.029**
	cnae(A.A)	$\rho_{6.2}$	-2.882	1.078	**0.007**
	cnae(C;R.A.M.)	$\rho_{6.2}$	-1.184	0.422	**0.005**
	n_contratos	ρ_8	-0.085	0.029	**0.004**
			$\boldsymbol{\tau}$		
	Intercepto	γ_0	16.420	6.455	**0.011**
	filiais	γ_7	0.457	0.218	**0.036**
	n_contratos	γ_8	-0.117	0.049	**0.017**
	log(renda)	γ_9	-2.49	0.907	**0.006**
Modelo 2			$\boldsymbol{\mu}$		
	Intercepto	β_0	0.126	1.258	0.92
	log(share)	β_2	-0.126	0.041	**0.002**
	eng(Possui)	β_4	0.836	0.248	**< 0.001**
	regiao(Nordeste)	$\beta_{5.2}$	-0.487	0.155	**0.001**
	regiao(Outras)	$\beta_{5.3}$	-0.026	0.097	0.784
	estudo	β_6	0.274	0.035	**< 0.001**
	log(renda)	β_7	-0.702	0.166	**< 0.001**
	cnae(I.T.)	$\beta_{8.1}$	0.537	0.148	**< 0.001**
	cnae(A.A)	$\beta_{8.2}$	0.922	0.181	**< 0.001**
	cnae(C;R.A.M.)	$\beta_{8.3}$	0.327	0.132	**0.014**
	n_contratos	β_{10}	$1.52*10^{-3}$	$7.31*10^{-4}$	**0.038**
			$\boldsymbol{\sigma}$		
	Intercepto	λ_0	-0.053	0.052	0.304
	n_contratos	λ_1	$-2.92*10^{-3}$	$8.14*10^{-4}$	**< 0.001**
			$\boldsymbol{\nu}$		
	Intercepto	ρ_0	4.328	1.298	**< 0.001**
	idade	ρ_1	-0.029	0.017	**0.098**
	rotatividade	ρ_2	-1.664	0.604	**0.006**
	estudo	ρ_5	-0.321	0.105	**0.002**
	cnae(I.T.)	$\rho_{6.1}$	-1.088	0.497	**0.029**
	cnae(A.A)	$\rho_{6.2}$	-2.881	1.078	**0.007**
	cnae(C;R.A.M.)	$\rho_{6.2}$	-1.184	0.422	**0.005**
	n_contratos	ρ_8	-0.085	0.029	**0.004**
			$\boldsymbol{\tau}$		
	Intercepto	γ_0	16.420	6.455	**0.011**
	filiais	γ_7	0.457	0.218	**0.036**
	n_contratos	γ_8	-0.117	0.049	**0.017**
	log(renda)	γ_9	-2.493	0.908	**0.006**

5.2 Análise de Diagnóstico

Definido o modelo, uma etapa importante é verificar sua adequabilidade. Vimos que para o modelo de regressão beta inflacionado em zero e um existem técnicas que, através do resíduo quantis aleatorizados, nos permitem detectar pontos influentes, constatar possíveis erros de escolha de função de ligação, examinar a distribuição do resíduos e até mesmo verificar se escolha do modelo está correta ou não.

Ao analisar a Figura 10, podemos ver como os resíduos estão distribuídos. Nota-se pela Figura 10 (a) que os resíduos estão distribuídos de maneira simétrica em torno do zero e a curva de sua densidade tem forma leptocúrtica. Pela Figura 10 (b) observamos que, mesmo com alguns poucos desvios na calda, há fortes indícios de normalidade dos resíduos, uma vez que existem poucos pontos fora da reta.



(a) (b)

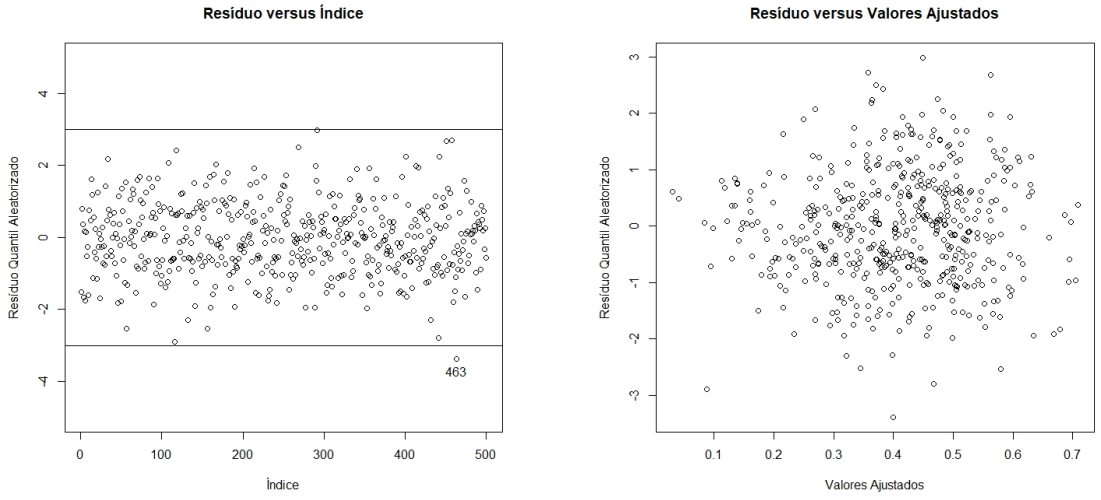
Figura 10: Densidade do Resíduo Quantis Aleatorizados (a) QQ Norm do Resíduos Quantis Aleatorizados (b).

A Tabela 15 traz um resumo dos resíduos. Os resultados apresentados nesta tabela reforçam o que apresentado pela Figura 10. Podemos perceber que a média dos resíduos é bem próxima de zero e que seu coeficiente de assimetria similarmente está próximo de zero. Observa-se também que o coeficiente de curtose está maior que 3, indicando que a densidade dos resíduos possui uma curva leptocúrtica. Por fim, o teste de Shapiro-Wilk, cuja a hipótese nula é dada por H_0 : Os resíduos possuem distribuição normal, resultou num p-valor de 0.8733, portanto não há evidencias para rejeitar que os resíduos seguem um distribuição normal.

Tabela 15: Resumo dos Resíduos.

Resumo dos Resíduos Quantis Aleatorizados.	
Média	-0.04
Variância	0.991
Coef. de assimetria	0.049
Coef. de Curtose	3.058
P-valor Shapiro-Wilk teste de Normalidade	0.8733

Observando a Figura 11 podemos detectar alguns pontos discrepantes e se há alguma forma funcional indicando erro de escolha de função de ligação. Como limiar para detectar a presença de pontos influentes foi escolhidos aqueles são maiores que 3 ou menores que -3, as escolha desses valores é devido ao fato da variável resposta possuir valores discretos e contínuos. Podemos notar pela Figura 11 (a) há uma observação que ultrapassa o limite, porém como ela não se afastou muito e tampouco a retirada trouxe mudanças significativas nos parâmetros, essa observação se manteve no modelo. Já na Figura 11 (b) podemos verificar que não há uma tendencia sistemática notável indicando algum erro de escolha de função de ligação.

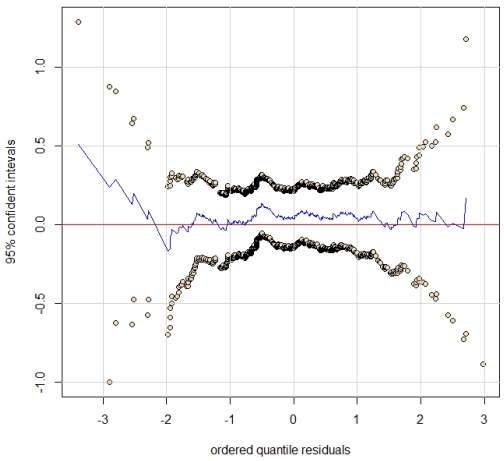


(a) (b)

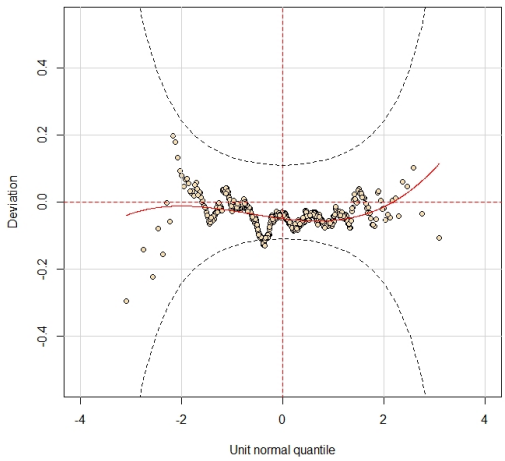
Figura 11: Resíduo Quantis Aleatorizados versus Índice (a) Resíduos Quantis Aleatorizados versus Valores Ajustados (b).

Na Figura 12 pode-se ver um ajuste geral do tanto que a escolha do modelo é adequada ou não. Pelo DTOP plot, podemos notar que não há nenhum indício de falta de ajuste do modelo. Pelo Worm plot notamos que alguns poucos pontos ultrapassaram o limite, porém não é nada que indique uma má qualidade do ajuste; o formato em S

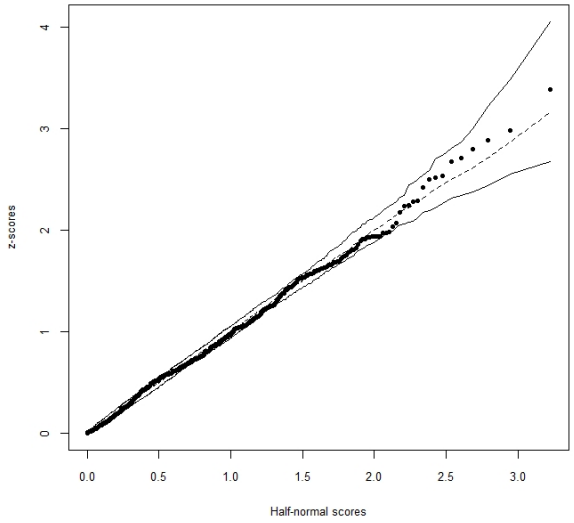
crescente nos indica que há caldas pesadas, entretanto esse formato em S não é muito acentuado, o que também indica que a escolha da distribuição é adequada. Já no gráfico de Envelope Simulado, percebemos que há um pequeno número de observações situadas no limite do envelope, enquanto as demais estão dentre dele, o que indica que não há indícios de afastamento da suposição de distribuição beta inflacionada em zero e um para a escolha da variável resposta.



(a)



(b)



(c)

Figura 12: DTOP Plot (a) Worm Plot (b) Gráfico de Envelope Simulado (c).

A título de informação, a Tabela 16 mostra as medidas de informação de todos os modelos apresentados nesse estudo. Podemos notar que o modelo escolhido detém o menor AIC e BIC entre eles. Em relação ao Pseudo R^2 o modelo escolhido é 0.62506, ou

seja, o modelo é capaz de explicar cerca de 62.506% das variações registradas na variável dependente. É sabido que quantos mais variáveis no modelo, maior será seu Pseudo R^2, porém comparado ao modelo com todas as variáveis não se teve uma perda tão grande e como nesse trabalho se trata de um modelo inferencial, essa pequena perda não deve ser considerada como um impasse.

Tabela 16: Medidas de Informação.

	Modelo c/ todas as variáveis	Modelo 1	Modelo 2
AIC	282.4254	254.7871	251.9295
BIC	493.1558	368.5815	357.2948
Pseudo R^2	0.66823	0.62763	0.62506

5.3 Interpretação do Modelo

Preliminarmente, é fundamental advertir que, ao interpretar uma variável do modelo, as demais serão mantidas constantes. Dito isso, visto que a análise de diagnósticos indicou um bom ajuste do modelo, ele pode ser interpretado assim como foi enunciado na Seção 4.1.

De antemão, podemos ver que algumas variáveis aparentam ser bastante importantes, como *estudo*, *renda*, *cnae* e *n_contratos* - umas vez que elas aparecem em duas equações do modelo, enquanto as demais apenas uma. A variável *n_contratos* era de se esperar, em virtude de que variáveis *offset* se adéquam bem nas equações de inflação.

O componente μ está relacionado a proporção de mulheres - visto que esta proporção esteja no intervalo contínuo (0,1). Desse modo, portanto, na interpretação desse componente, considere excluído a possibilidade de assumir proporção 0 ou 1. Logo, podemos inferir pela Tabela 14, que caso ela possua engenheiros, a chance de ter uma proporção de mulheres maior aumenta em 83.6% comparado com uma que não possui. A chance de ter uma proporção de mulheres maior diminui 48.7% se a empresa está situada na região Nordeste e 2.6% nas demais contrastado com a região Sudeste. Conseguimos notar também que a chance de ter uma proporção de mulheres maior é diretamente proporcional ao tempo de estudo médio; se considerarmos, por exemplo, que um curso de graduação duram em média 5 anos, podemos inferir que a chance de ter uma proporção de mulheres maior aumenta em 137% numa empresa com 5 anos a mais de tempo de estudo médio comparado com a que não tem. Nota-se também a chance de ter uma proporção de mulheres maior é inversamente proporcional à renda média da empresa e ao *Market Share*. Por fim, defrontado com as demais cnae's, a chance de ter uma proporção de mulheres maior aumenta 53.7% se a atividade econômica da empresa for Industria de Transformação,

92.2% Alojamento e Alimentação e 32.7% Comércio; Reparação de Veículos Automotores e Motocicletas. A Figura 13 mostra como cada preditor está relacionado à $\log(\mu/(1-\mu))$ acompanhado dos erros padrões.

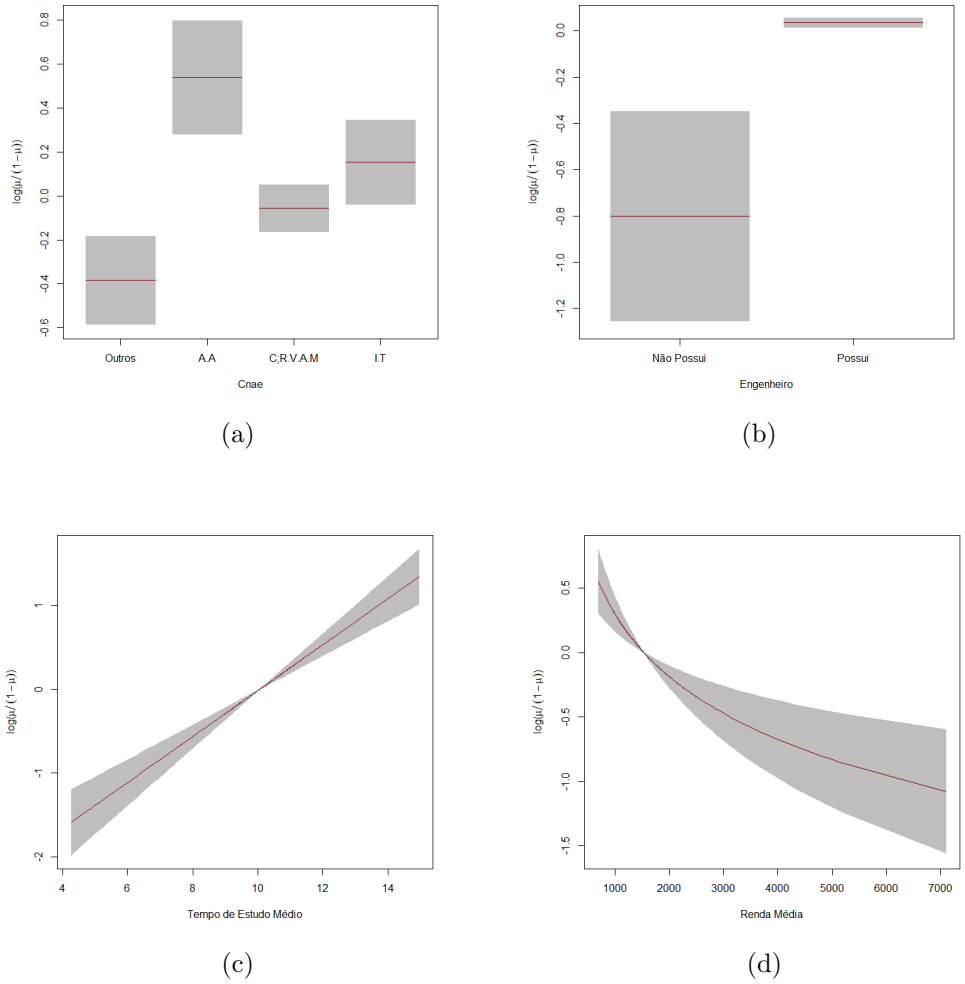
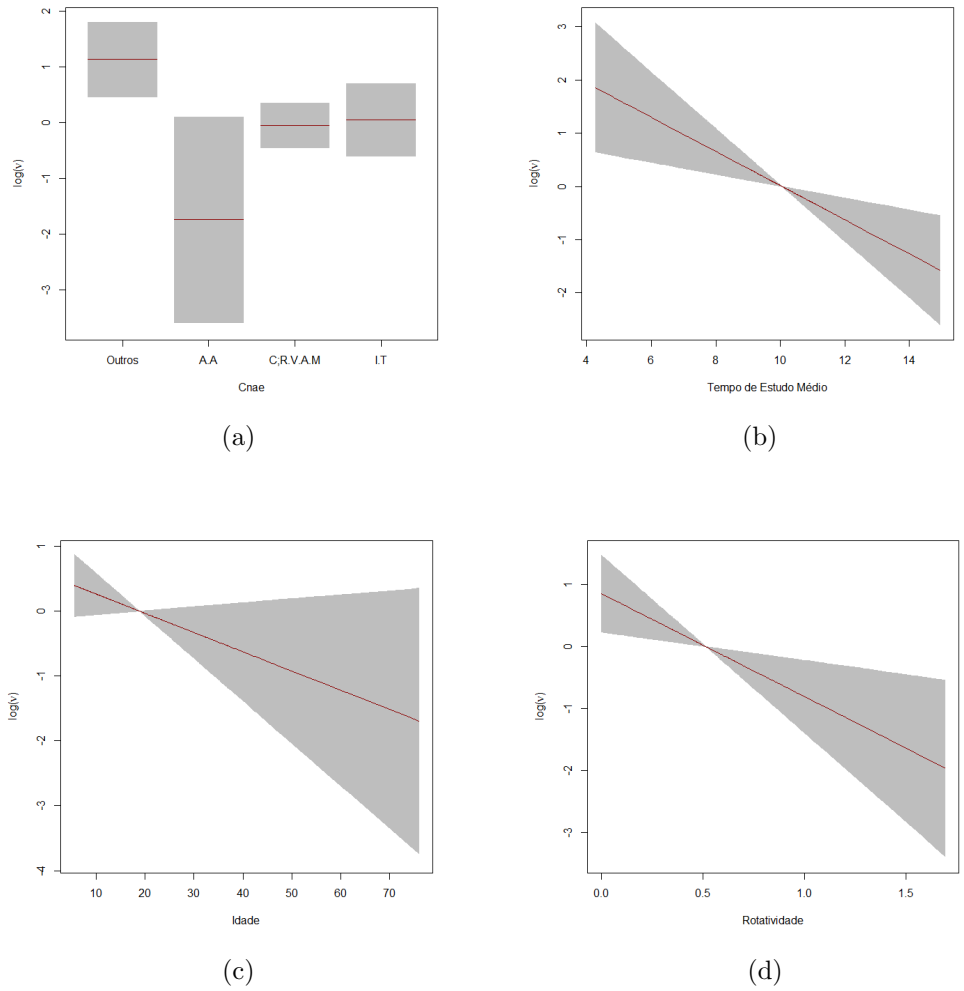


Figura 13: Cnae (a) Engenheiro (b) Tempo de estudo médio (c) Renda Média (d).

No tocante, ao componente ν, relacionado a chance da empresa possuir uma proporção de mulher igual a 0, todas a variáveis que entraram no modelo se deram inversamente proporcional a chance de não ter nenhuma mulher, exceto o intercepto. Podemos inferir que, no caso de uma empresa i ser 5 anos mais velha que a j, a chance da empresa i não ter nenhuma mulher é 14.5% menor em relação a j. Também temos que quanto maior taxa de admissão e demissão da empresa menor a chance dela não ter nenhuma mulher. De maneira análoga, temos que a chance da empresa não possuir nenhuma mulher diminui em 32.1% aumento de um ano no tempo de estudo médio. Enfim, comparado com as demais cnae's, a chance da empresa não ter nenhuma mulher diminui 108.8% se a atividade econômica da empresa for Industria de Transformação, 288.1% Alojamento e

Alimentação e 118.4% Comércio; Reparação de Veículos Automotores e Motocicletas. A Figura 14 expressa o modo como cada variável está relacionada à $\log(\nu)$ acompanho dos erros padrões.



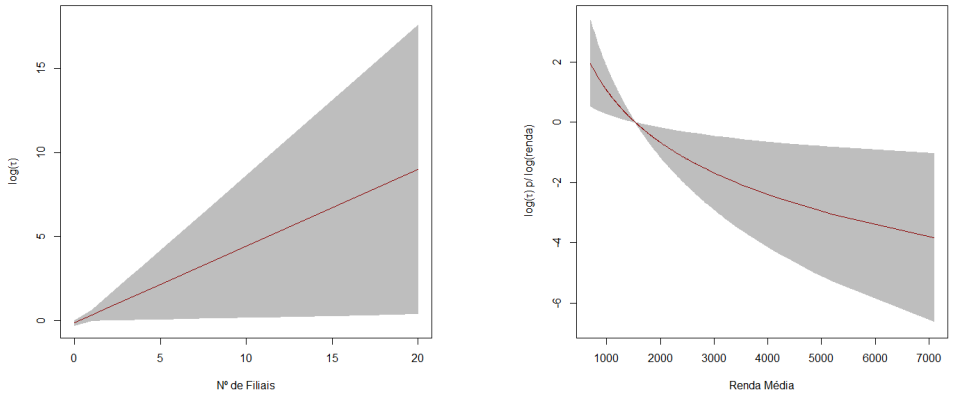
(a) (b)

(c) (d)

Figura 14: Cnae (a) Tempo de estudo médio (b) Idade da Empresa (c) Taxa de Rotatividade (d).

Com relação ao componente τ, relacionado a chance da empresa possuir uma proporção de mulher igual a 1, todas a variáveis que entraram no modelo se deram inversamente proporcional a chance de só ter mulher, exceto o número de filiais que a empresa possui. Sendo assim, podemos inferir que, para cada filial a mais que a empresa tem, a chance da mesma só ter mulher aumenta em 45.7%. Este resultado é contra-intuitivo, uma vez que espera-se que conforme aumente o número de filiais, diminui o a chance de só ter mulher na empresa; porém, vale ressaltar que esse resultado pode ser dado pelo fato de que, nas empresas que só há mulher na amostra, exitem algumas delas que possuem um número "alto" de filiais. Constatamos também que, para cada aumento de um ponto

percentual na renda média da empresa, a chance dela só ter mulher diminui em 249.3%. Por fim, inferimos que, quanto maior a participação da empresa no mercado, menor a chance dela só possuir mulher. A Figura 15 expressa o modo como cada variável está relacionada à $\log(\tau)$ acompanho dos erros padrões.



(a) (b)

Figura 15: Quantidade de Filiais (a) Renda Média da Empresa (b) Market Share (c).

A interpretação da variável nº de contratos é a mais modesta comparado as demais, quanto maior o nº de contratados na empresa menor a chance dela só possuir mulher ou de não ter nenhuma. Assim como o nº de contratos ser inversamente proporcional aos parâmetros do intervalo contínuo (0,1), μ e σ.

6 Considerações Finais

O presente trabalho teve como objetivo apresentar o modelo de regressão beta inflacionado em zero e um, no qual é indicado para modelar problemas em que o domínio da variável é um intervalo [0,1], ou até mesmo [a,b], b>a. Assumimos para variável resposta a distribuição beta inflacionada em zero e um - que corresponde a mistura da distribuição beta com a Bernoulli. Nesse contexto, esse modelo se mostra bastante flexível e com capacidade de interpretação mais rica comparado aos modelos usuais. Este trabalho foi desenvolvido computacionalmente com auxilio do pacote gamlss do *software* R, no qual está implementado a estimação dos parâmetros assim como as técnicas de diagnóstico do modelo.

Na conjuntura social brasileira, foi utilizado como aplicação a proporção de mulheres nas empresas. Percebemos que, de acordo com os dados de 2016 obtidos em parceria com o IPEA, 8.8% das empresas pertencentes na amostra não possuíam nenhuma mulher e 4.8% só haviam mulheres. Através desses dados, o objetivo era propor um modelo inferencial para verificar quais fatores estão influenciando na proporção de mulheres da empresas, tanto no intervalo contínuo (0,1) quanto modelar os fatores que influenciam a empresa a só ter mulheres ou não ter nenhuma.

Na Seção 5, vimos que, de maneira geral, o modelo de regressão beta inflacionado em zero e um se ajustou bem os dados. Além disso, conseguimos tirar algumas conclusões importantes, como, por exemplo, quanto maior o tempo de estudo médio da empresa, maior a proporção de mulheres nas empresas e que, quanto maior a renda média da empresa, menor a proporção de mulheres nas empresas.

A obtenção dos dados para esse trabalho não foi obtida de maneira ideal. Observamos pela Tabela 6 que a maioria das empresas pertencentes na amostra estão situadas na região Sudeste. De fato, pelo cenário econômico brasileiro atual, era de se esperar que essa região se destacasse em relação as demais, porém vimos no modelo final que, diferente da região Nordeste, as demais não se mostraram significativas no modelo. Logo, como perspectiva futura, será dado um foco maior no delineamento amostral de modo que, as empresas fiquem melhores alocadas nas regiões.

Vimos na Seção 3.2 que um dos pressupostos do modelo é que a variáveis explicativas sejam exógenas e sabemos que a existência de variáveis endógenas pode ser problemática na estimativas dos parâmetros. É fácil encontrar na literatura diversos estudos que lidam sobre a relação da escolaridade com uma media de renda como, por exemplo, (MINCER, 1974). Os estudo empíricos dessa área mostram o problema de endogeneidade da variável tempo de estudo, uma vez que ela sozinha não consegue medir os retornos da educação na variável de interesse. Esses estudos sugerem alguns usos de variáveis instrumentais para

contornar esse problema, como QI, escolaridade do pai e da mãe e habilidade. Entretanto, no nosso estudo haviam uma limitação de variáveis disponíveis para serem trabalhadas. Portanto, temos como perspectiva futura, a obtenção de variáveis instrumentais e a implementação do método de máxima verossimilhança de informação limitada para o modelo de regressão beta inflacionado em zero e um.

Referências

AKAIKE, H. A new look at the statistical model identification. In: *Selected Papers of Hirotugu Akaike*. [S.l.]: Springer, 1974. p. 215–222.

ATKINSON, A. C. *Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis*. [S.l.], 1985.

BOLFARINE, H.; SANDOVAL, M. C. *Introdução à inferência estatística*. [S.l.]: SBM, 2001. v. 2.

BRUSCHINI, C.; LOMBARDI, M. R. O trabalho da mulher brasileira nos primeiros anos da década de noventa. *Anais*, p. 483–516, 2016.

BUUREN, S. v.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001.

DJENNAD, A. et al. Detrended transformed owen's plot a diagnostic tool for checking the adequacy of a fitted model distribution. 2012.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004.

GELMAN, A. et al. *Bayesian data analysis*. [S.l.]: Chapman and Hall/CRC, 2013.

GUJARATI, D. N.; PORTER, D. C. *Econometria Básica-5*. [S.l.]: Amgh Editora, 2011.

HIRATA, H. S. *Nova divisão sexual do trabalho?: um olhar voltado para a empresa e a sociedade*. [S.l.]: Boitempo, 2002.

JOSÉ, I. J. S. P. I.; FURLAN, D. C. *MÉTODOS NUMÉRICOS*. Tese (Doutorado) — Universidade Federal do Paraná, 2006.

JUNIOR, O. A. G. *Aplicação de Modelos Beta Inflacionados de Zeros para Análise de Dados Longitudinais*. Dissertação (Mestrado) — Universidade Estadual de Maringá, 2017.

LOPES, G. D. C. Uma análise da eficiência dos municípios paraibanos em relação a utilização dos recursos do programa bolsa família com a utilização da análise envoltória de dados e da regressão beta inflacionada. *Universidade Federal da Paraíba*, 2017.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. [S.l.]: CRC press, 1989. v. 37.

MCLACHLAN, G.; KRISHNAN, T. *The EM algorithm and extensions*. [S.l.]: John Wiley & Sons, 2007. v. 382.

MCLACHLAN, G.; PEEL, D. *Finite Mixture Models*. [S.l.]: Wiley, 2004. (Wiley Series in Probability and Statistics). ISBN 9780471654063.

ME. *Relação Anual de Informações Sociais*. 2015. Ministério da Economia. Acesso em 3 mar. 2019. Disponível em: ⟨http://trabalho.gov.br/rais⟩.

MINCER, J. Schooling, experience, and earnings. human behavior & social institutions no. 2. ERIC, 1974.

MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. Half-normal plots and overdispersed models in r: The hnp package. *J. Stat. Softw*, v. 81, p. 23, 2017.

NAGELKERKE, N. J. et al. A note on a general definition of the coefficient of determination. *Biometrika*, Oxford University Press, v. 78, n. 3, p. 691–692, 1991.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, v. 135, n. 3, p. 370–384, 1972.

OSPINA, M. R. *Modelos de regressão beta inflacionados*. Tese (Doutorado) — Universidade de São Paulo, 2008.

OWEN, A. B. Nonparametric likelihood confidence bands for a distribution function. *Journal of the American Statistical Association*, Taylor & Francis, v. 90, n. 430, p. 516–521, 1995.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2004.

PEREIRA, G. H. d. A. *Modelos de regressão beta inflacionados truncados*. Tese (Doutorado) — Universidade de São Paulo, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: ⟨https://www.R-project.org/⟩.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C*, v. 54, n. 3, p. 507–554, 2005. Disponível em: ⟨https://EconPapers.repec.org/RePEc:bla:jorssc:v:54:y:2005:i:3:p:507-554⟩.

SAS Institute Inc. *SAS 9.4 Logon Manager*. Cary, NC, USA, 2007. Disponível em: ⟨https://odamid.oda.sas.com⟩.

SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978.

SEDLACEK, G. L.; SANTOS, E. C. A mulher cônjuge no mercado de trabalho como estratégia de geração da renda familiar. Instituto de Pesquisa Econômica Aplicada (Ipea), 1991.

SIMAS, A. B.; BARRETO-SOUZA, W.; ROCHA, A. V. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, Elsevier, v. 54, n. 2, p. 348–366, 2010.

STASINOPOULOS, D. et al. Flexible regression and smoothing: The gamlss packages in r. *GAMLSS for Statistical Modelling. GAMLSS for Statistical Modeling*, 2015.

STASINOPOULOS, D. M.; RIGBY, R. A. et al. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, v. 23, n. 7, p. 1–46, 2007.

STASINOPOULOS, M.; RIGBY, B. *Generalised Additive Models for Location Scale and Shape*. [S.l.], 2018. Disponível em: ⟨http://www.gamlss.org/⟩.

Anexos

A.1 Códigos Utilizados

```
##Importação dos dados
source("C:/Users/Jean/Desktop/tcc_modelo/dados.R")

##Estabelecendo as categorias de referencia
dados.mod$regiao2 <- relevel(dados.mod$regiao2, "Sudeste")
dados.mod$cnae2 <- relevel(dados.mod$cnae2, "Outros")


##Selecionando o modelo
library(gamlss)
set.seed(1)

##### MODELO SATURADO #######
mod.saturado <- gamlss(prop ~ idade+log(share) +rotatividade +
   eng1 +
regiao + estudo + log(renda) + filiais + cnae2 + n_contratos,
nu.formula  = ~ n_contratos +  idade + rotatividade + eng1 +
regiao + estudo  + filiais + cnae2 + log(renda) + log(share) ,
tau.formula = ~ n_contratos + idade + rotatividade + eng1 +
regiao + estudo  + filiais + cnae2  + log(renda) + log(share),
sigma.formula =~ n_contratos,
family = BEINF,  data =dados.mod)
summary(mod.saturado)


##### ESCOLHENDO MODELO #######
mod.6 <- gamlss(prop ~  log(share) + eng1 +  regiao + estudo +
   log(renda) + cnae2 + n_contratos,
nu.formula  = ~ idade + rotatividade + estudo + cnae2 + n_
   contratos,
tau.formula = ~ filiais + n_contratos +log(renda),
sigma.formula =~ n_contratos,
family = BEINF,  data =dados.mod)
summary(mod.6)
```

```
#MODELO 1
mod.saturado1 <- gamlss(prop ~  log(share) + eng1 +  regiao +
   estudo + log(renda) + cnae2 + n_contratos,
nu.formula  = ~ idade + rotatividade + estudo + cnae2 + n_
   contratos,
tau.formula = ~ filiais + n_contratos +log(renda),
sigma.formula =~ n_contratos,
family = BEINF,  data =dados.mod)
summary(mod.saturado1)

#MODELO 2
mod.saturado2 <- gamlss(prop ~  log(share) + eng1 +  regiao2 +
   estudo + log(renda) + cnae2 + n_contratos,
nu.formula  = ~  idade + rotatividade + estudo + cnae2 + n_
   contratos,
tau.formula = ~ filiais + n_contratos +log(renda),
sigma.formula =~ n_contratos,
family = BEINF,  data =dados.mod)
summary(mod.saturado2)


###ANLISE DE DIAGNOSTICO

plot(mod.saturado2)

res <- mod.saturado2$residuals
plot(density(res), xlab = "Resíduo Quantil Aleatorizado", ylab =
   "Densidade",
main = "Densidade Estimada")

qqnorm(res, xlab = "Quantis Teóricos", ylab = "Quantis Amostrais"
   )
qqline(res)
shapiro.test(res)

plot(res, ylim = c(-5,5), xlab = "Índice", ylab = "Resíduo
   Quantil Aleatorizado",
main = "Resíduo versus Índice")
abline(-3,0)
abline(3,0)
```

```
identify(res)

ajustado <- fitted(mod.saturado2)
plot(ajustado,res, xlab = "Valores Ajustados", ylab = "Resíduo
   Quantil Aleatorizado",
main = "Resíduo versus Valores Ajustados")


#WORM PLOT
wp(mod.saturado2)

dtop(mod.saturado2)
#ENVOLOPE
d.fun <- function(obj) resid(obj)
s.fun <- function(n, obj) rBEINF(n, obj$mu.fv, obj$sigma.fv, obj$
   nu.fv, obj$tau.fv)
f.fun <- function(y.){
gamlss(y. ~ log(share) + eng1 +  regiao2 + estudo + log(renda) +
   cnae2 + n_contratos,
nu.formula  = ~  idade + rotatividade + estudo + cnae2 + n_
   contratos,
tau.formula = ~ filiais + n_contratos +log(renda),
sigma.formula =~ n_contratos,
family = BEINF,  data =dados.mod)
}

library(hnp)
hnp(mod.saturado2, newclass = TRUE, diagfun = d.fun, simfun = s.
   fun,
fitfun = f.fun, xlab = "Half-normal scores", ylab = "z-scores",
main = "", pch = 20, cex = 1, cex.lab = .8, cex.axis = .8)

### MEDIAS E CRITERIOS DE INFORMAÇÃO
AIC(mod.saturado2)
BIC(mod.saturado2)
Rsq(mod.saturado2, type = "both")


### GRAFICOS DE INTERPRETAÇÃO
term.plot(mod.saturado2, terms = 1, , xlab = "Share" , ylab =
   expression(paste("log(",mu/(1-mu), ")")))
```

```
term.plot(mod.saturado2, terms = 2, xlab = "Engenheiro" , ylab =
   expression(paste("log(",mu/(1-mu), ")")))
term.plot(mod.saturado2, terms = 3, xlab = "Região" , ylab =
   expression(paste("log(",mu/(1-mu), ")")))
term.plot(mod.saturado2, terms = 4, xlab = "Tempo de Estudo Médio
   " , ylab =expression(paste("log(",mu/(1-mu), ")")))
term.plot(mod.saturado2, terms = 5, xlab = "Renda Média" , ylab =
   expression(paste("log(",mu/(1-mu), ")")))
term.plot(mod.saturado2, terms = 6, xlab = "Cnae" , ylab =
   expression(paste("log(",mu/(1-mu), ")")))
term.plot(mod.saturado2, terms = 7, xlab = "N° de Contratos" ,
   ylab =expression(paste("log(",mu/(1-mu), ")")))

term.plot(mod.saturado2, what = "nu" , terms = 1, xlab = "Idade"
,ylab =expression(paste("log(",nu, ")")), se= F )
term.plot(mod.saturado2, what = "nu" , terms = 2, xlab = "
   Rotatividade"
,ylab =expression(paste("log(",nu, ")")))
term.plot(mod.saturado2, what = "nu" , terms = 3, xlab = "Tempo
   de Estudo Médio"
,ylab =expression(paste("log(",nu, ")")))
term.plot(mod.saturado2, what = "nu" , terms = 4, xlab = "Cnae"
,ylab =expression(paste("log(",nu, ")")))
term.plot(mod.saturado2, what = "nu" , terms = 5, xlab = "N° de
   Contratos"
,ylab =expression(paste("log(",nu, ")")))



term.plot(mod.saturado2, what = "tau" , terms = 1, xlab = "N° de
   Filiais"
,ylab =expression(paste("log(",tau, ")")))
term.plot(mod.saturado2, what = "tau" , terms = 3, xlab = "Renda
   Média"
,ylab =expression(paste("log(",tau, ") p/ log(renda)")))
term.plot(mod.saturado2, what = "tau" , terms = 2, xlab = "N° de
   Contratos"
,ylab =expression(paste("log(",tau, ")")))
```